Breath selection methods for compact mass spectrometry breath analysis

Breath selection methods for compact mass spectrometry breath analysis.

K Rosenthal[1,3], DM Ruszkiewicz[2], H Allen[1], MR Lindley[1,3], MA Turner[2,3], & E Hunsicker[4].

[1] School of Sport, Exercise & Health Sciences, Loughborough University, Loughborough, UK
[2] Department of Chemistry, Loughborough University, Loughborough, UK
[3] Translational Chemical Biology Research Group
[4] Department of Mathematical Sciences, Loughborough University, Loughborough, UK

E-mail: m.r.lindley@lboro.ac.uk

Received xxxxxx
Accepted for publication xxxxxx
Published xxxxxx

Abstract

Compact mass spectrometry (CMS) is a versatile and transportable analytical instrument that has the potential to be used in clinical settings to quickly and non-invasively detect a wide range of relevant conditions from breath samples. The purpose of this study is to optimise data preprocessing protocols by three proposed methods of breath sampling, using the CMS. It also lays out a general framework for which data processing methods can be evaluated. **Methods** This paper considers data from three previous studies, each using a different breath sampling method. These include a peppermint washout study using continuous breath sampling with a purified air source, an exercise study using continuous breath sampling with an ambient air source, and a single breath sampling study with an ambient air source. For each dataset, different breath selection (data preprocessing) methods were compared and benchmarked according to predictive performance on a validation set and quantitative reliability of m/z bin intensity measurements. **Results** For both continuous methods, the best breath selection method improved the predictive model compared to no preselection, as measured by the 95% CI range for Youden's index, from 0.68-0.86 to 0.86-0.97 for the exercise study and 0.69-0.82 to 1.00-1.00 for the peppermint study. The reliability of intensity measurements for both datasets (as measured by median relative standard deviation), was improved slightly by the best selection method compared to no preselection, from 18% to 14% for the exercise study and 7% to 5% for the peppermint study. For the single breath samples, all the models resulted in perfect prediction, with a 95% CI range for Youden's index of 1.00-1.00. The reliability of the proposed method was 38%. **Conclusion** The method of selecting exhaled breath from CMS data can affect the reliability of the measurement and the ability to distinguish between breath samples taken under different conditions. The application of appropriate data processing methods can improve the quality of the data and results obtained from CMS. The methods presented will enable untargeted analysis of breath VOCs using CMS to be performed.

Keywords: exhaled breath, data preprocessing, mass spectrometry, modelling

---

1. Introduction

Metabolites are low-weight molecular compounds which are produced by a biological system during metabolic processes and may have an endogenic or exogenic source. Metabolomics aims to identify all the metabolites in a biological system and produce a 'fingerprint' of compounds within or released by a substance (Liland, 2011). The more metabolites in a biological system that can be identified, the more complete an understanding of that system can be. While the general aim of metabolomics is to identify every metabolite within a biological system, the analysis of part of a biological system, such as urine or blood, can give crucial information about the organism while limiting the analysis necessary.

Breath analysis is particularly attractive as it is non-invasive (Poli et al., 2010). The analysis of volatile organic compounds (VOCs) in breath has been shown to detect differences between healthy participants and patients with various diseases including chronic obstructive pulmonary disease (Van Berkel et al., 2010), oral cancer (Bouza et al., 2017) and lung cancer (Wehinger et al., 2007). However, there is yet to be widespread application of metabolomics methods as diagnostic tools due to the limitations of the current methods. For example, gas-chromatography mass-spectrometry (GC-MS) is expensive, requires specialist operators, and complex analyte mixtures, such as breath, may require long running times, in some cases of around 60 min (Perez-Hurtado et al., 2017). This makes it difficult to use in a clinical setting where there is a high throughput of samples and a requirement to keep costs low.

Compact mass spectrometry (CMS) is comparatively inexpensive and smaller than many mass spectrometry instruments, such that the instrument can be transported. VOCs can be introduced to ambient ionisation sources, such as atmospheric pressure chemical ionisation (APCI) and electrospray ionisation (ESI), which interface with the CMS directly, thus removing the requirement for sample preparation (Perez-Hurtado et al., 2017). The technique demonstrated promising results for real-time analysis of VOCs in breath providing almost instantaneous results (Heaney et al., 2016). However, due to the novelty of the technique in the breath field, a standardised CMS data processing protocol does not currently exist. This paper aims to provide such protocol.

Breath sample data collected using the CMS will have many sources of noise, particularly as continuous breath samples can include contaminants from the ambient air. Thus, some method is required at either the sampling stage or the data preprocessing stage to separate datapoints in the raw MS data that can be attributed to breath from those that can be attributed to ambient air. Some previous studies using different mass spectrometry based methods have used a valve system at the sampling stage based on CO_2 concentration to divert unwanted breath; therefore, only collecting the end-tidal breath (King et al., 2009; Pizzini et al., 2018). In others, a 'breath tracker' algorithm has been used at the data preprocessing stage to differentiate between alveolar and inspired phases in continuous real time proton-transfer-reaction mass-spectrometry (PTR-MS) monitoring, based on a selected indicator mass which is expected to be considerably higher in breath than in ambient air, usually acetone (Trefz et al., 2013). Continuous breath samples have been shown to increase total ion count at the onset of exhalation and reduce it during inhalation, with the peaks in total intensity corresponding to the pressure trace in the breathing mask (Heaney et al., 2016). Therefore, changes in the total intensity over time could be used to select the data corresponding to the exhalation. This may be more reliable than using a signal indicator mass, as it will be less sensitive to noise.

Ambient air will contaminate the sample, even if the alveolar phase is selected accurately, as any compounds inhaled from the environment will be exhaled for some time after, particularly if they are hydrophobic or in large quantities; therefore, even the use of a clean air source cannot completely remove this problem (Turner, 2016). However, requiring participants to inhale filtered air for a minimum period prior to providing the sample will reduce some of the exogenous compounds, and will introduce only minimal contamination during the sampling. This method is useful for resting samples. However, a continuous flow rate is not suitable for the deep and quick breathing patterns typical during exercise; therefore, providing the participant with enough air through a filtered source requires a sophisticated system, and thus an ambient air source is often used instead.

This study aims to compare data preprocessing methods for selecting exhaled breath data from continuous and single breath samples, to allow for measured differences to be detected between conditions or sample type. The purpose is not to identify the VOCs causing these differences. However, a discussion of the potential identity relating to previously identified VOCs is included, as a check on the data preprocessing methods. The breath selection methods were compared using three measures: reliability of resulting intensity measurements, number of potential biomarkers identified, and the ability of biomarker intensity measurements to separate samples from participants under two conditions. Multivariate data analysis was used to identify potential biomarkers of exercise and potential biomarkers of peppermint oil ingestion in continuous breath samples, and to develop models to distinguish between resting and exercise breath samples, and between baseline and 120 minutes post-ingestion of peppermint oil. Similar methods were used to create models to distinguish single breath samples from ambient air.

2. Methods

2.1 Ethics

All participants gave informed consent and completed a health screen questionnaire. The study was approved by the Loughborough University ethics committee and followed the guidelines of the Declaration of Helsinki for Human Research.

2.2 Data collection

All the exhaled breath samples were analysed using the Advion expression compact mass spectrometer (Advion, NY, USA). The instrument was operated in positive ion mode for all experiments. All spectra were acquired over a range between m/z 30-300. For both continuous sampling methods conditions of the CMS ion source parameters were: capillary temperature 250 °C, capillary voltage 42 V, source voltage 15 V, APCI gas temperature 40 °C, corona discharge voltage 5 kV, with a scan rate of 300 ms/scan.

For the single breath study, conditions of the CMS ion source parameters were: capillary temperature 225 °C, capillary voltage 90 V, source voltage 15 V, APCI source gas temperature 350°C, corona discharge 4 uA. The scan rate was 700 ms/scan.

2.2.1 Continuous real-time sampling of exercise dataset

Five healthy, recreationally trained, male participants (age 18-35 years) breathed through a mouthpiece with a valve allowing ambient air to be inhaled, producing a total sample set of seven as two participants completed the protocol twice on two separate days. The final dataset consisted of seven resting breath samples, seven exercise breath samples, and seven ambient air samples. A one-way non-return valve connected directly to a heated transfer line transported air from the mask to the APCI source of the CMS at a constant rate for 1 min.

Samples were taken under two conditions: resting and during the last stage of a maximal incremental exercise test. The participant was seated for approximately 20 mins prior to taking the resting sample. The maximal exercise test was conducted on a Monark Ergomedic 874e basket-loaded friction-braked cycle ergometer (Monark Exercise AB, Vansbro, SWE). The participants were asked to maintain a cadence of 90 rpm throughout the duration of the test. A 10-minute warm-up at 135 W preceded the maximal test, which began at 180 W and increased by 45 W every three minutes. Exhaled air samples were taken during the last minute of each increment. The participant was asked to inform the researcher when they felt they would be able to maintain a 90 rpm cadence for only one further minute, at which point a final exhaled breath sample was taken regardless of the time point within the stage.

The participant's final breath sample was used for further analysis. As participants finished at different stages, the final samples were taken during a workload of between 360-540 W, 13-27 min after the onset of exercise. A sample of ambient air was also taken.

2.2.2 Continuous real-time sampling of peppermint dataset

The peppermint study used a second breath sampling method using a mask, with filtered air provided at a constant rate of 20 L/min. Sixteen healthy participants (seven males and nine females, age 22-53 years) provided samples at baseline and 60, 120, 240, and 360 min post-ingestion of a commercially available 200 mg peppermint oil capsule (Boots, Nottingham, UK). These methods are described in detail by Heaney et al. (2016).

The ability to distinguish between differing breath samples was a test of the data processing methods, which required a dataset with clear differences. Initial examination of the data indicated that breath samples were consistently different from baseline at 120 min post-ingestion of peppermint oil, across participants; therefore, that time point was chosen as the testing sample. As indicated by Heaney et al. (2016), some participants were late responders. Selecting 120 min over 60 min post-ingestion should limit the effect of the delayed response on distinguishing between breath samples pre- and post-ingestion.

2.2.3 Single breath study dataset

Five breath samples were collected from one participant on four different days (total of 20 samples), all under resting conditions. The single breath sample was collected via offline disposable Haldane tube breath sampler, adapted from GAS Dortmund sampling approach, which captured the end tidal volume of a single breath. This sample was then connected directly to the CMS transfer line, allowing breath to be analysed. This method will be described in more detail in a future paper.

The instrument measured the room air for around 10-30 seconds before the CMS transfer line was attached to the sampler. This allowed for a direct comparison between the ambient air and the breath sample without the instrument being reset.

2.3 Data processing

Mass spectra data were exported from the Advion data express software, via OpenChrom (Wenig & Odermatt, 2010), to CSV files suitable for input into R. The OpenChrom software binned the intensities into windows of 1 m/z (therefore, bin 35 was the sum of all intensities for $34.5 \geq m/z < 35.5$). The resolution of the instrument is

1 m/z; therefore, this step does not reduce the resolution of the results. All subsequent data processing and analyses were performed using R 3.4.1 (R Core Team, 2013).

Figure 1 shows a workflow of the steps used to create models to distinguish between resting and exercise breath samples, and between baseline and post-ingestion of peppermint oil. There were no ambient air samples available for the peppermint study, therefore the first variable selection step was not applied.

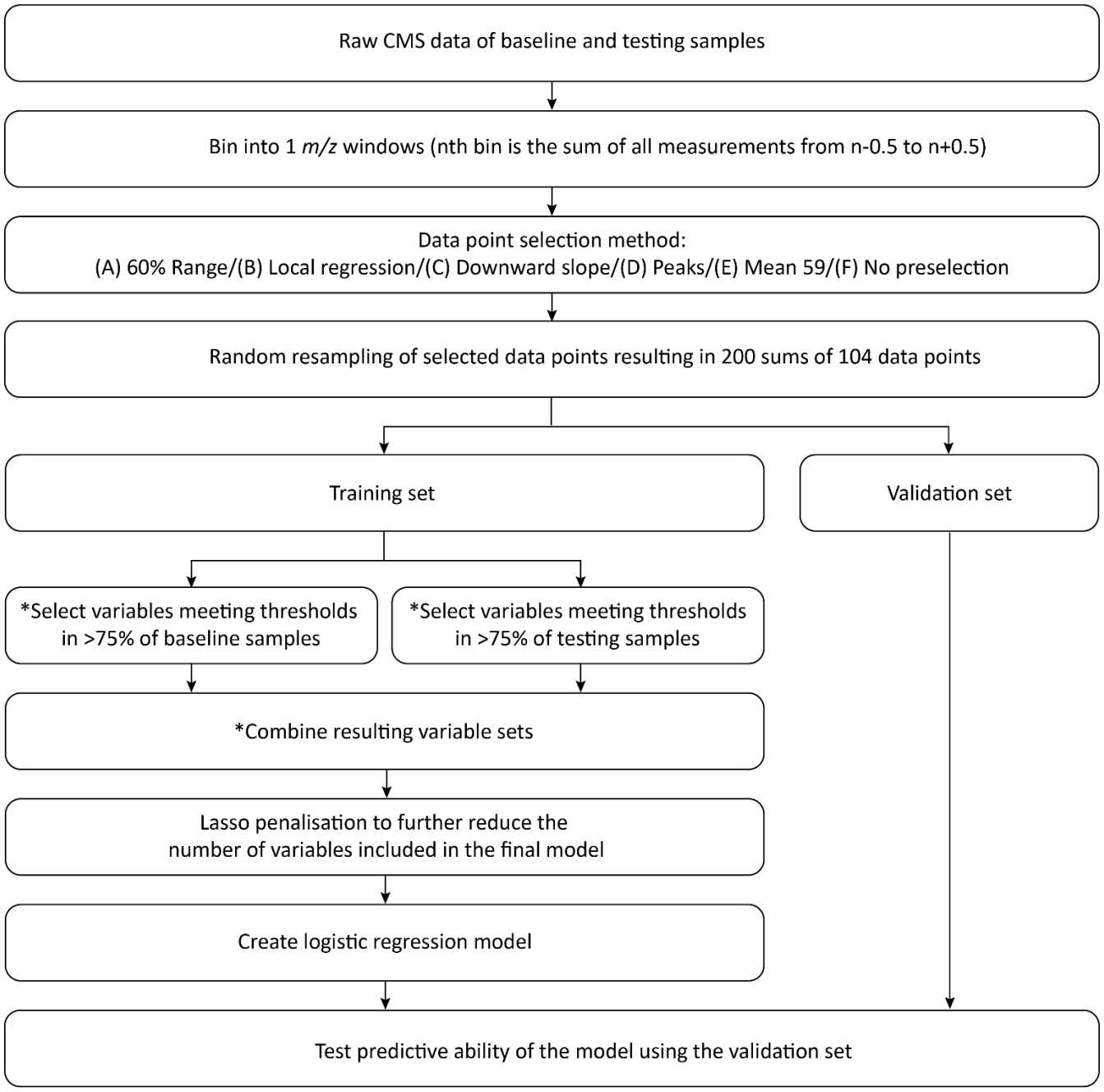

Figure 1. *A workflow to select exhaled breath data from continuous samples and create models to distinguish between conditions. The workflow was repeated to apply each of the different datapoint selection methods to each dataset. Thus, six models were created for each dataset. An asterisk (*) indicates steps that were only applied to the exercise study data and refers to the thresholds of a signal-to-noise ratio above three and a fold-change above two, for breath versus air.*

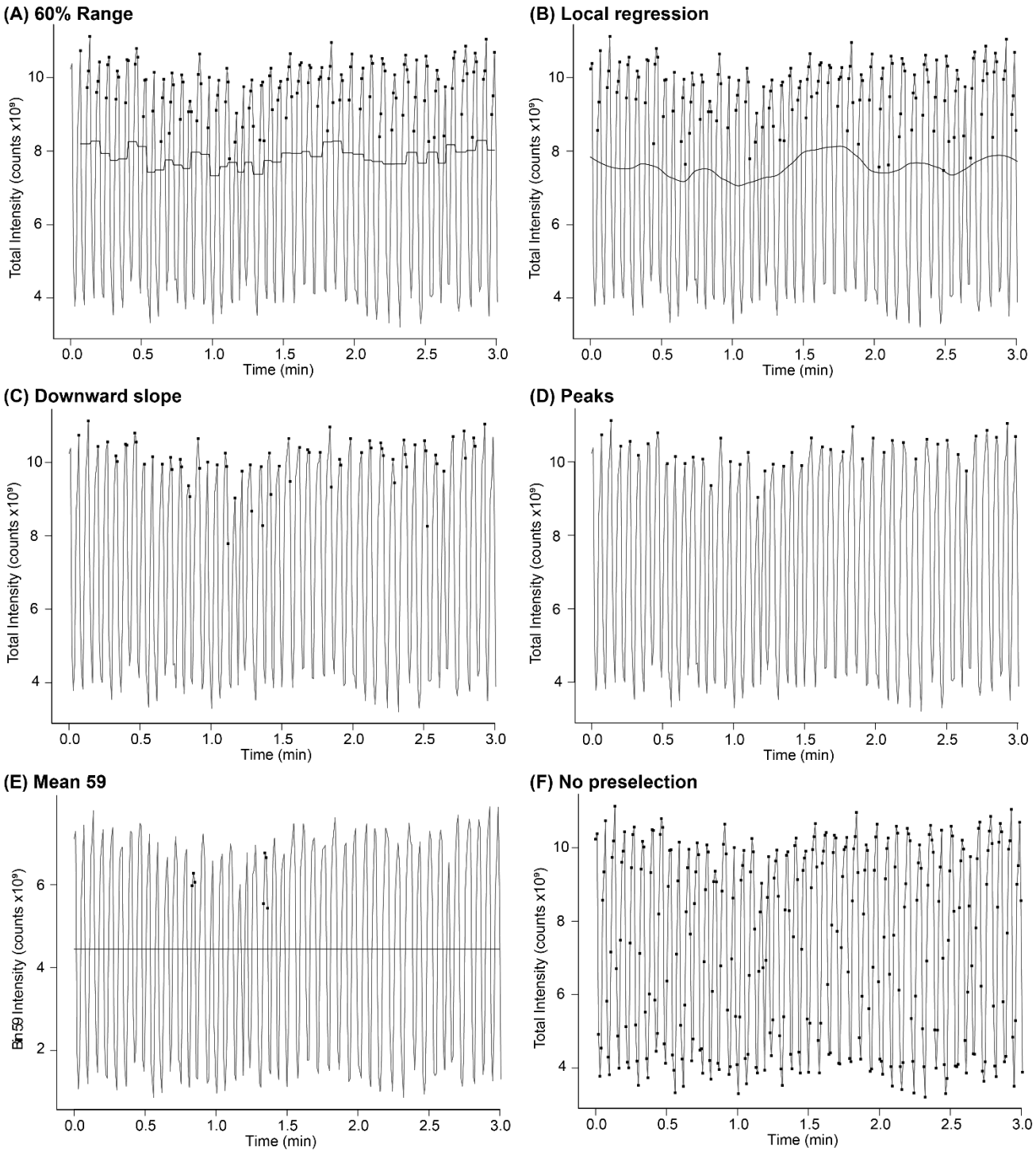

Figure 2. *Graphs showing the data points selected using the six datapoint selection methods, for one participant's baseline data. The grey line is the total intensity or bin 59 intensity graph, the black dots are the selected data points and, where appropriate, the lower threshold is shown by a black line. A time point was selected if: (A) the total intensity was above 60% of the range between peaks; 60% range, (B) the total intensity was above the local regression line plus the standard deviation of the local regression line; Local regression, (C) the total intensity was above 60% of the range of each downward slope; Downward slope, (D) the peaks in total intensity; Peaks, (E) the intensity of bin 59 (acetone) greater than the bin 59 mean of the whole sample, and with a gradient of less than 2.5% between the first and last points of each breath; Mean 59, (F) all data points selected; No preselection.*

2.3.1 Continuous samples data preprocessing methods

Five datapoint selection methods were tested, with the aim to exclude data that was not measuring the exhaled breath. Four methods were based on the total intensity: (A) 60% Range, (B) Local regression, (C) Downward slope, and (D) Peaks. One method was based on the intensity of bin 59 (acetone a known breath marker); (E) Mean 59. The method using bin 59 was based on the description by Schwoebel et al. (2011), where data points above the mean and with a breath gradient of <2.5% were selected. The corresponding time points from the whole spectra were selected for each method. The whole sample with no preselection was also tested; (F) No preselection. Graphs and further details to demonstrate these methods are shown in Figure 2.

The datapoint selection methods were applied to the resting samples (baseline) and the samples from the last stage of the maximal exercise test (testing) from the exercise study, and to the baseline samples (baseline) and samples 120 min post-ingestion of peppermint oil (testing) from the peppermint study. Following the datapoint selection (or no preselection), 200 resamples were created by summing 104 randomly resampled time points (one-minute worth), for each participant's baseline, testing and ambient air data. This method was used to reduce the influence of technical variance due to machine drift by sampling across it.

2.3.2 Single breath data preprocessing methods

All zeros in the data were removed. A threshold for each bin was calculated as the mean plus three times the standard deviation of the first 10 scans (of ambient air). Five methods of summarising the breath data for each bin were compared; (1) the sum of the first 10 scans, (2) the sum of the data points above the threshold in the first 10 scans, (3) the sum of all scans above the threshold, (4) a stable selection method, and (5) no selection.

The stable selection method included more complex data selection algorithms, where a value was calculated for each bin as either:

- the sum of all data points until five consecutive points are below the threshold, with maximum cut off point at 1 min after the first breath scan

Or

- the sum of the first 10 scans, if any of the following conditions were met:
    - the first five consecutive points below the threshold were within the first 10 scans
    - the first point above the threshold was after 30 s
    - no points were above the threshold

Additionally, all the air measurements until two before the first breath measurement were used to calculate the threshold and, if there were still fewer than two non-zero scans, the standard deviation of the whole sample for that bin was used. Figure 3 demonstrates how methods 1-4 select data. For (5) no selection, all data points after the first increase in total intensity were summed.

A measure of intensity for each bin for the corresponding ambient air sample was calculated by randomly resampling from the first 10 points of each sample. The number of data points resampled was equivalent to the number of data points summed to calculate the measure of intensity for breath for that bin, to ensure a direct comparison was possible.

2.4 Statistical analyses

2.4.1 Creation of training and validation subsets

The data were split into training and validation subsets. For the exercise study, there were five samples in the training set (four participants with one repeat) and two samples in the validation set (two participants with one repeat from training set). The repeats were from different testing sessions and were therefore considered separately. However, the above parameters were chosen to ensure that the model was not based on only three participants and the validation set did not contain only repeats. For the peppermint study, there were twelve samples in the training set and four samples in the validation set. For the single breath study, there were five samples in the validation set, from across the four days of sampling, and 15 samples in the training set.

The training and validation sets were formed by random selection of samples for validation, under the parameters defined above. The same training and validation sets were used to evaluate each exhaled breath selection method. Therefore, the main results of whether the breath selection methods improved the separation, would not be affected by the presence of repeated participants.

2.4.2 Variable selection

Potential biomarkers were selected using the criteria of a signal-to-noise ratio over three and a fold-change over two. The signal-to-noise ratio was defined as the difference in means between a breath sample and the ambient air sample, divided by the standard deviation of the ambient air sample, which was taken to represent the noise. The fold-change was simply the mean of a breath sample divided by the mean of the ambient air sample. As ambient air data was not available for the peppermint study, this step was only applied to the exercise and single breath studies.

For the exercise study, the following criteria were applied to each participant:

$$(\bar{x}_B - \bar{x}_A)/\sigma_A > 3 \;\&\; \bar{x}_B/\bar{x}_A > 2$$

$$(\bar{x}_T - \bar{x}_A)/\sigma_A > 3 \;\&\; \bar{x}_T/\bar{x}_A > 2$$

Where $\bar{x}_B$ = mean of the baseline resamples, $\bar{x}_T$ = mean of the testing resamples, $\bar{x}_A$ = mean of the ambient air resamples, and σ_A = standard deviation of the ambient air resamples. Bins that met the criteria for one or both of the conditions in >75% of the participants were kept.

For the single breath study, a bin was considered a potential biomarker if it met the following criteria on at least three of the four days:

$$(\bar{x}_B - \bar{x}_A)/\sigma_A > 3 \;\&\; \bar{x}_B/\bar{x}_A > 2$$

Where $\bar{x}_B$ = mean of the breath samples for that day, $\bar{x}_A$ = mean of the ambient air samples for that day, and σ_A = standard deviation of the ambient air samples for that day. Bins that met the criteria for in three or more days were kept.

Lasso penalisation was applied to all datasets to reduce some of the coefficients to zero based on high intercorrelation, with the remaining variables being kept for the final model. This was applied using the glmnet R package function 'cv.glmnet' (Friedman, Hastie, & Tibshirani, 2010), with lambda optimised by 10-fold cross-validation and selected as the largest value such that error is within one standard error of the minimum. This method automatically applies standardisation (centring and z-scaling), which prevents features with large variance from being penalised less.

The lasso penalisation method was selected to evaluate the data processing methods while avoiding issues such as overfitting and over-selection of predictor variables common in least-squares based methods (McNeish, 2015). However, this could result in relevant biomarkers being excluded from the final model. This method is designed to find the clearest differences between the samples, not identify all the potential biomarkers. Therefore, it may not be appropriate for other metabolomics approaches.

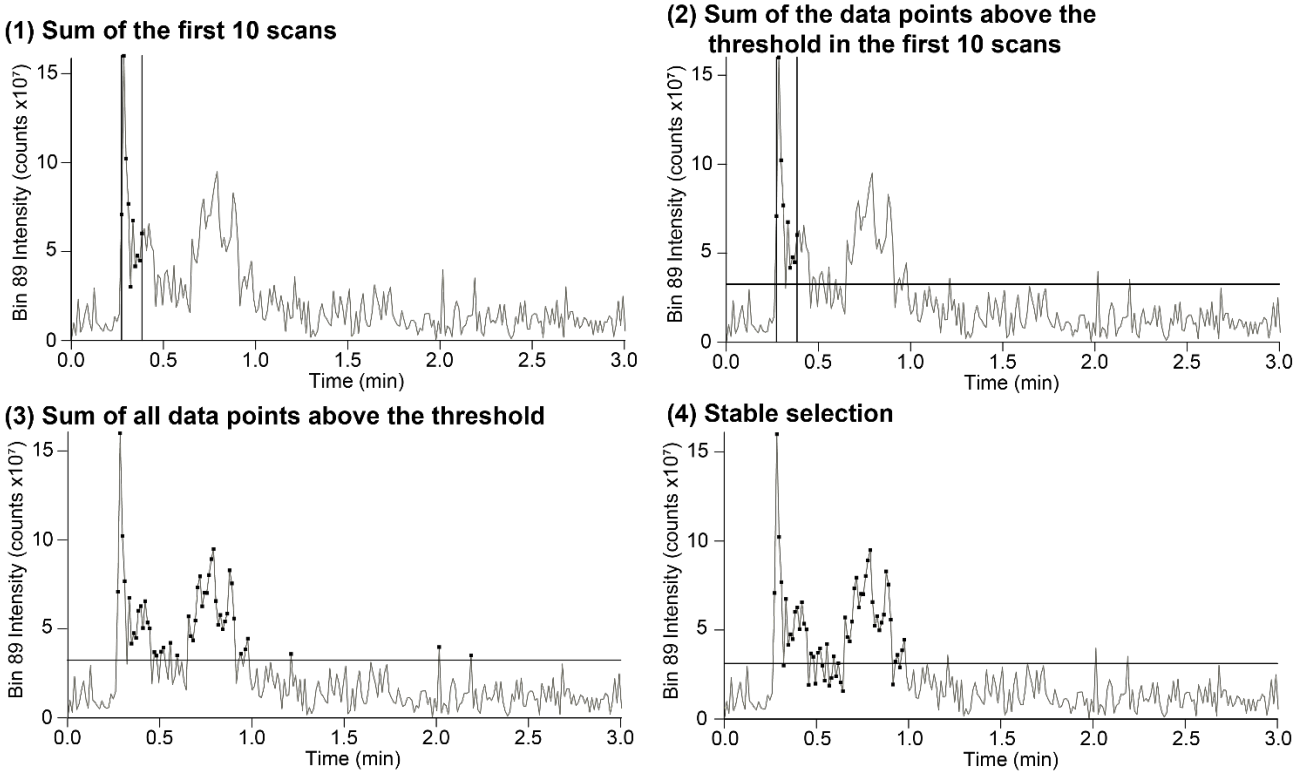

Figure 3. *Four methods to select the exhaled breath data from the single breath samples; (1) the sum of the first 10 scans, (2) the sum of the data points above the threshold in the first 10 scans, (3) the sum of all scans above the threshold, and (4) the stable selection method. The grey line represents the intensity of bin 89, the vertical black lines indicate the time of the first and tenth scans of breath, the horizontal line is the threshold (the mean plus three times the standard deviation of the first ten scans of ambient air), and the black dots indicate points that were selected by the method. This sample and bin were chosen to show an example of where all the selection methods would select different datapoints.*

2.4.3 Model creation and evaluation

For each datapoint selection method, a binary logistic regression model was built using the training data from the variables selected by the lasso penalisation, to separate baseline from testing samples or breath from ambient air samples. The glm function in R was used, with no scaling applied as this does not affect the predictive ability of the model. The ability of the model to distinguish between the samples was assessed using the validation set. The optimal model for each datapoint selection method was selected using the maximum of Youden's index: J = sensitivity + specificity – 1 (Bewick, Cheek, & Ball, 2004). The final models were compared using the same statistic, with a greater value indicating a greater prediction ability, and using the residual deviance, with a lower value indicating less variance left unexplained. A 95% confidence interval for Youden's index was created using 2000 stratified bootstrap replicates. Single-marker models were made for each bin included in the best model of each dataset, to further examine predictive ability of individual bins.

2.4.4 Reliability of resting breath measurements

For each m/z bin (30-300), an intensity was measured over several time points; the reliability is defined in this context in terms of the relative standard deviation (RSD) of the measurements at individual time points. The reliability of baseline breath markers was compared among the different datapoint selection methods and no preselection. For the continuous methods, the RSD for each of the variables was calculated as:

$$RSD\ (\%) = 100\ \times (\frac{standard\ deviation}{mean\ \times\ \sqrt{104}})$$

The standard deviation and mean were calculated from the individual data points after breath selection, not the random resamples; therefore, dividing by $\sqrt{104}$ corrected for the random resampling of 104 datapoints, as actual analytical replicates were not available. Therefore, this measure of reliability incorporates biological variance, variance among breaths and the stages of each breath, and analytical variance.

For the single breath study, RSD was calculated after applying a breath selection method, for each day as:

$$RSD\ (\%) = 100\ \times (\frac{standard\ deviation}{mean})$$

The RSD was compared across datapoint selection methods in resting samples for all the studies, using the median RSD for potential resting breath biomarkers.

3. Results

Apart from (F) No preselection, (B) Local regression selected the most data points for both continuous breath sampling methods (Table 1). (E) Mean 59 selected the fewest data points, and, for two (out of 14) exercise study samples and four (out of 32) peppermint study samples, it did not select any data points. If no data points were selected, both samples for that participant were removed from the modelling analysis.

3.1 Models

3.1.1 Exercise study models

The (A) 60% Range model had the highest optimal Youden's index for the exercise data; however, the 95% confidence interval was large (Table 2). The (D) Peaks model had a high Youden's index with a small confidence interval and a small residual deviance (Table 2), and thus was chosen as the best model. The (D) Peaks method was also able to identify the most potential breath markers in comparison to ambient air (20). Three of the single-marker models resulted in a Youden's index 95% CI ≥0.5 for both the training and validation sets. The difference in measured intensity between conditions for those bins is shown in Figure 4(a).

3.1.2 Peppermint study models

For the peppermint models, three of the models had perfect separation (Youden's index J = 1, Table 4). Of the three models with perfect separation, (D) Peaks had the lowest residual deviance, and thus was chosen as the best model, although (A) 60% Range, (B) Local regression, and (C) Downward slope all produced good models. As there were no ambient air samples available, lasso penalisation was applied to all 271 bins. Three of the single-marker models resulted in a Youden's index 95% CI ≥0.5 for both the training and validation sets. The difference in measured intensity between conditions for those bins is shown in Figure 4(b).

Table 1. *The mean percentage of data points selected by each method.*

Breath sampling method	(A) 60% Range	(B) Local regression	(C) Down-ward slope	(D) Peaks	(E) Mean 59	(F) No preselection
Room air	29.9 ± 2.7	35.8 ± 6.8	24.9 ± 2.6	17.7 ± 2.7	3.3 ± 2.9	100 ± 0
Filtered air	43.1 ± 7.6	46.9 ± 6.8	22.3 ± 4.8	13.5 ± 3.8	3.6 ± 1.9	100 ± 0

Table 2. *Models to distinguish between exercise and resting breath samples.*

Breath selection method	Number of baseline markers	Number of testing markers	Total markers meeting criteria	Total markers after lasso penalisation	Youden's index (95% confidence interval)	Residual deviance
(A) 60% Range	11	7	15	10	0.998 (0.326-1)	1.1×10^{-7}
(B) Local regression	11	7	15	12	0.960 (0.607-0.978)	1.1×10^{-7}
(C) Downward slope	9	6	11	9	0.917 (0.812-0.960)	135.7
***(D) Peaks**	18	7	20	10	0.925 (0.855-0.965)	5.2×10^{-8}
(E) Mean 59	10	4	13	5	0.50 (0.363-0.550)	1.3×10^{-8}
(F) No preselection	9	7	12	12	0.808 (0.677-0.863)	271.6

** The best model, chosen using the Youden's index and residual deviance.*

Table 3. *Models to distinguish between breath samples taken 120 min post-ingestion of peppermint oil and baseline breath samples.*

Breath selection method	Total markers after lasso penalisation	Youden's index (95% confidence interval)	Residual deviance
(A) 60% Range	35	1.00 (NA)	8.8×10^{-8}
(B) Local regression	36	0.998 (0.991-1.00)	9.1×10^{-8}
(C) Downward slope	29	1.00 (NA)	7.4×10^{-8}
***(D) Peaks**	27	1.00 (NA)	7.0×10^{-8}
(E) Mean 59	16	0.669 (0.566-0.714)	2.9×10^{-8}
(F) No preselection	28	0.770 (0.691-0.819)	1.0×10^{-7}

** The best model, chosen using the Youden's index and residual deviance.*

3.1.3 Single breath study models

All the models to distinguish between ambient air and breath had perfect separation, with a Youden's index of 1.00, regardless of datapoint selection method with little difference in residual deviance (Table 6). Methods 1-4 selected between 41 and 69 potential breath markers in comparison to ambient air. Method (5) No selection identified the least potential breath markers (17). Three of the single-marker models resulted in a Youden's index 95% CI ≥0.5 for both the training and validation sets. The difference in measured intensity between ambient air and breath for those bins is shown in Figure 4(c).

Table 4. *Models to distinguish ambient air and breath.*

Breath selection method	Total markers meeting criteria	Total markers after lasso penalisation	Youden's index (95% confidence interval)	Residual deviance
(1) Sum of first 10	41	5	1.00 (NA)	3.9×10^{-10}
(2) Sum of points above threshold in first 10	69	4	1.00 (NA)	4.6×10^{-10}
(3) Sum of all above threshold	55	11	1.00 (NA)	2.2×10^{-10}
(4) Stable selection	51	5	1.00 (NA)	3.4×10^{-10}
(5) No selection	17	3	1.00 (NA)	1.3×10^{-9}

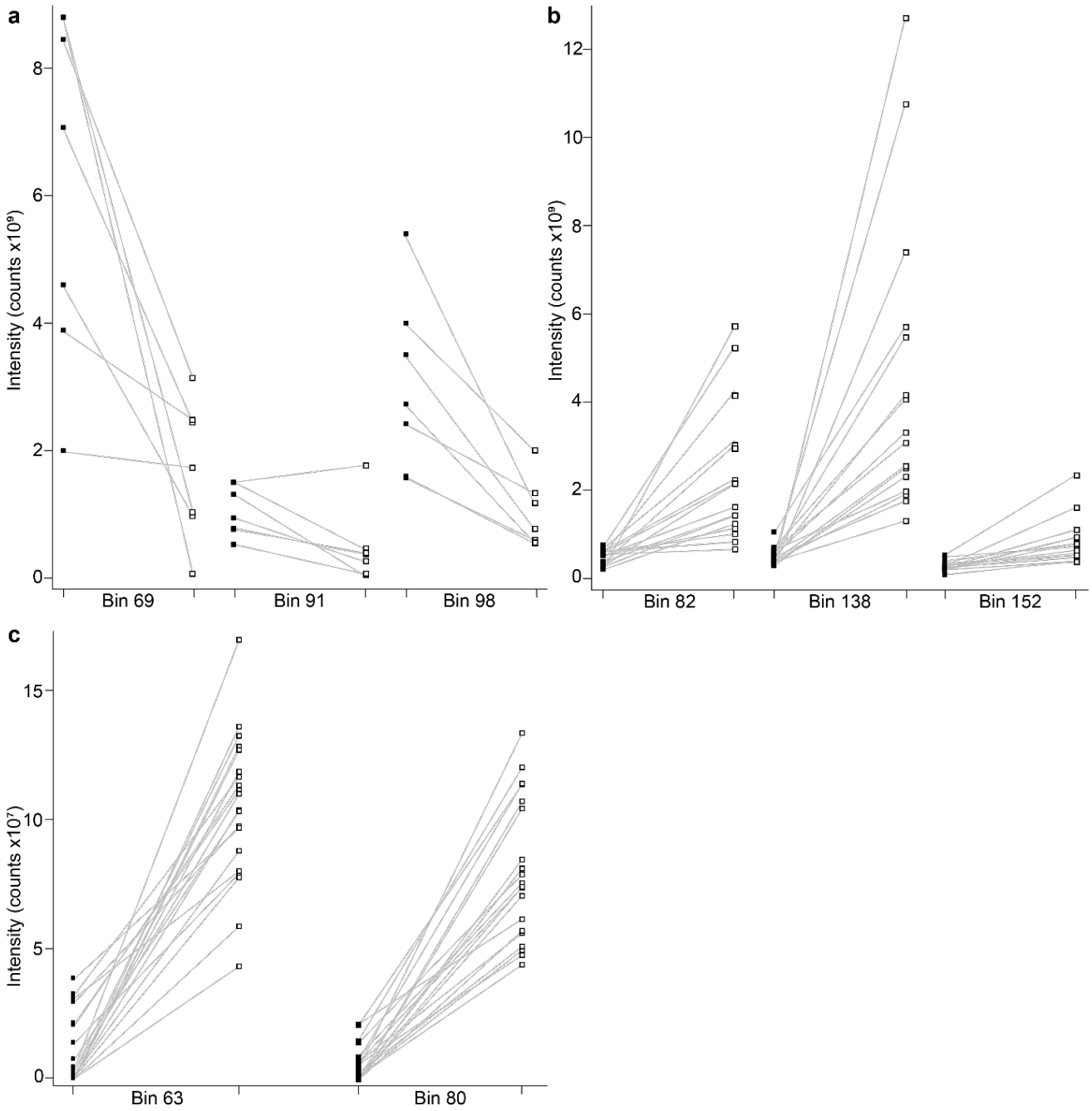

Figure 4. *The mean of the random resamples using the (D) Peaks method for each participant in the baseline (black squares) and testing (white squares with black outline) breath samples, for bins able to discriminate between: (a) rest and exercise breath samples, (b) baseline and post-ingestion of peppermint oil breath samples. (c) The breath (white squares with black outline) and ambient air (black squares) measurement calculated by the (4) Stable selection method for resting single breath samples and corresponding ambient air samples.*

3.1.4 Models - supplementary material

Appendix 1 shows which variables were selected by the signal-to-noise ratio and fold change criteria, and which were included in the final regression models, for all datapoint selection methods. Appendix 2 shows the relationship between RSD and signal-to-noise ratio for all bins included in the best final model. As lasso penalisation removes some variables based on high correlations, bins selected by the signal-to-noise ratio and fold change criteria, but not included in the final model, could potentially still be distinguishing biomarkers between breath samples taken under different conditions, or breath samples and ambient air. Correlation matrices are in Appendix 3. Tables showing the full results of the single-marker models are in Appendix 4. Appendix 5 shows an example of raw spectra for each of the sampling methods.

3.2 Reliability

There was little difference in the RSD among the continuous datapoint selection methods, with marginally worse reliability for (F) No preselection for both continuous breath sampling methods (Figure 5). For the ambient air method (exercise study), the median RSD for methods A-F were 14%, 15%, 15%, 14%, 16% and 18%, respectively. For the filtered air method (peppermint study), methods A-E had a median RSD of 5%, while (F) No preselection had a median RSD of 7%.

In the single breath study, the median RSDs for (2) Sum of points above the threshold in first 10 and (3) Sum of all above threshold were 60% and 77%, respectively, considerably greater than the other methods, with median RSDs of 31%, 38% and 27% for (1) Sum of first 10, (4) Stable selection method and (5) No selection, respectively (Figure 5(c)).

Figure 6 shows graphs of various bins to demonstrate how the different single breath selection methods affect the final sum of intensity for that bin, with some not reaching their maximum until after the tenth scan, and thus the final sum of intensity would not be representative for (1) Sum of the first 10 or (2) Sum of points above the threshold in first 10.

4. Discussion

This study compared various exhaled breath selection methods and developed models to determine the ability of the resulting data to discriminate between breath samples from participants under different conditions, or between breath and ambient air. Additionally, the reliability of the individual data points selected by each method was compared.

Although identification of biomarkers of peppermint ingestion/exercise was not the purpose of this study, it is useful to relate these findings to previous in the literature both as a basic check on the methods and as a way to help contextualise the methodological findings. However, as all three studies have very small sample sizes, the usefulness of the conclusions about these biomarkers in isolation is limited.

4.1 Continuous samples

4.1.1 Exercise versus resting models

The (A) Peaks method detected the most potential biomarkers for the exercise study and resulted in a particularly good predictive model of exercise when validated on a separate test set of samples. It also had the narrowest Youden's index 95% confidence interval and only overlapped with the (F) No preselection 95% confidence interval by 0.01. Generally, the datapoint selection methods resulted in a greater Youden's index and smaller residual deviance than (F) No preselection, suggesting that the selection methods were successful in removing noise from the ambient air and selecting the exhaled breath.

The exception to this was (E) Mean 59, which resulted in a particularly low Youden's index. However, given that the method selected very few data points (and in some cases none), it is possible that not enough data points were available to sample across the noise. Additionally, (E) Mean 59 selected fewer potential biomarkers, and therefore it is possible that ions crucial for determining the differences between resting and exercise breath may not have been selected, and thus not included in the final model. However, it did select the three bins identified as strong individual predictors (bins 69, 91 and 98). This method was adapted from Schwoebel et al. (2011), who used PTR-MS, with a scan every 0.19 s. This allowed for multiple scans during the alveolar phase and therefore a gradient during that phase could be calculated. In this study, it appears there was generally only one scan during that phase, and thus a gradient could not be calculated, and the data points would not be selected by this method.

The (C) Downward slope method resulted in a far greater residual deviance than the other datapoint selection methods for the exercise models. The method was designed to avoid selecting dead space air; however, it is possible that this actually resulted in not selecting data points containing useful information, probably from the alveolar phase. The model did not include bin 91, whereas the other datapoint selection methods did, which may indicate that this is a key marker for distinguishing between rest and exercise breath samples. Nevertheless, it was still somewhat better than (F) No preselection.

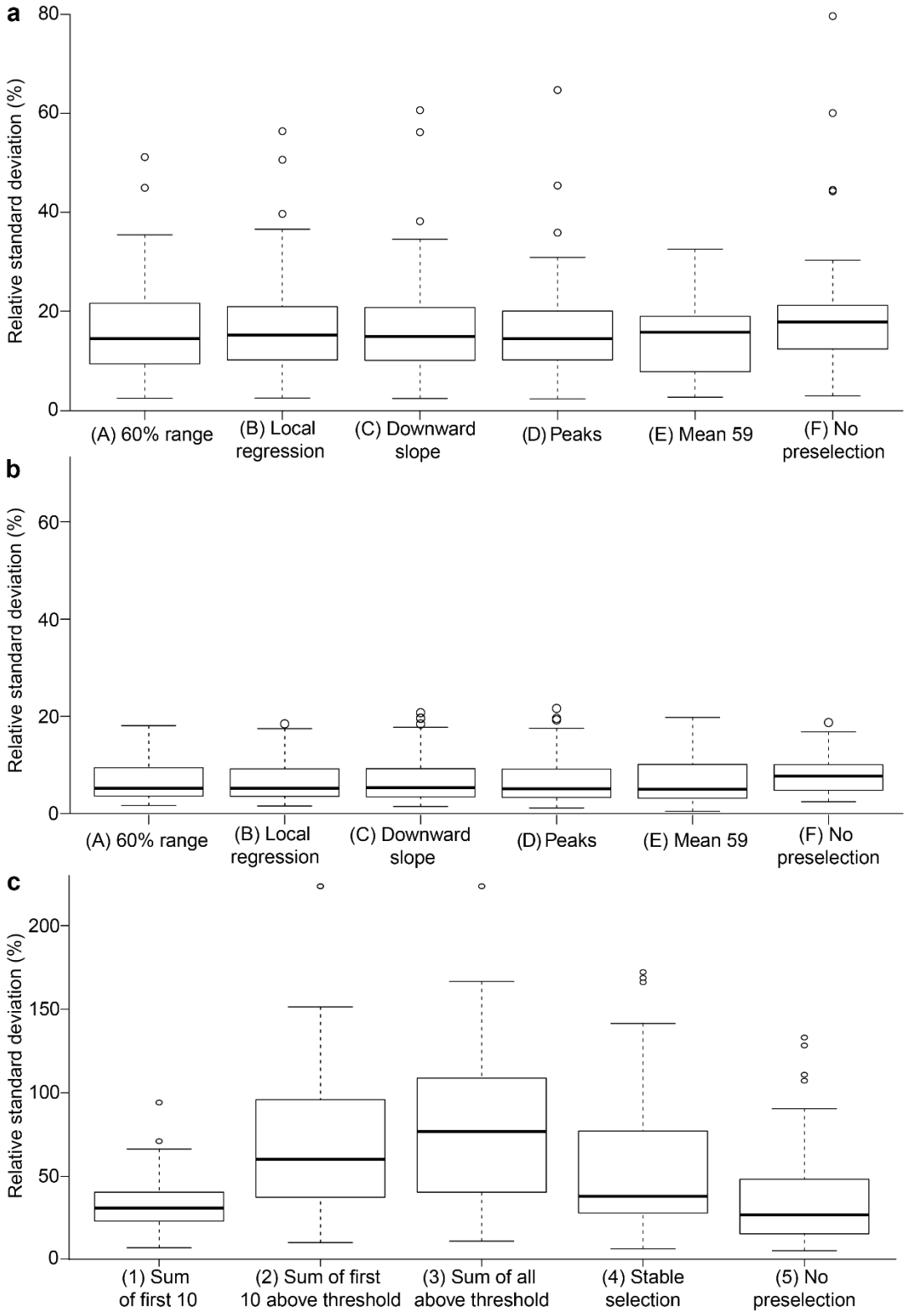

Figure 5. *Boxplots of the relative standard deviations (RSD) incorporating biological and analytical variability for baseline samples from all participants, for bins identified by all datapoint selection methods as potential breath biomarkers compared to ambient air. For: (a) continuous resting breath samples taken while inhaling ambient air, (b) continuous resting breath samples taken while inhaling filtered air, and (c) single breath samples. The open circles refer to outliers.*

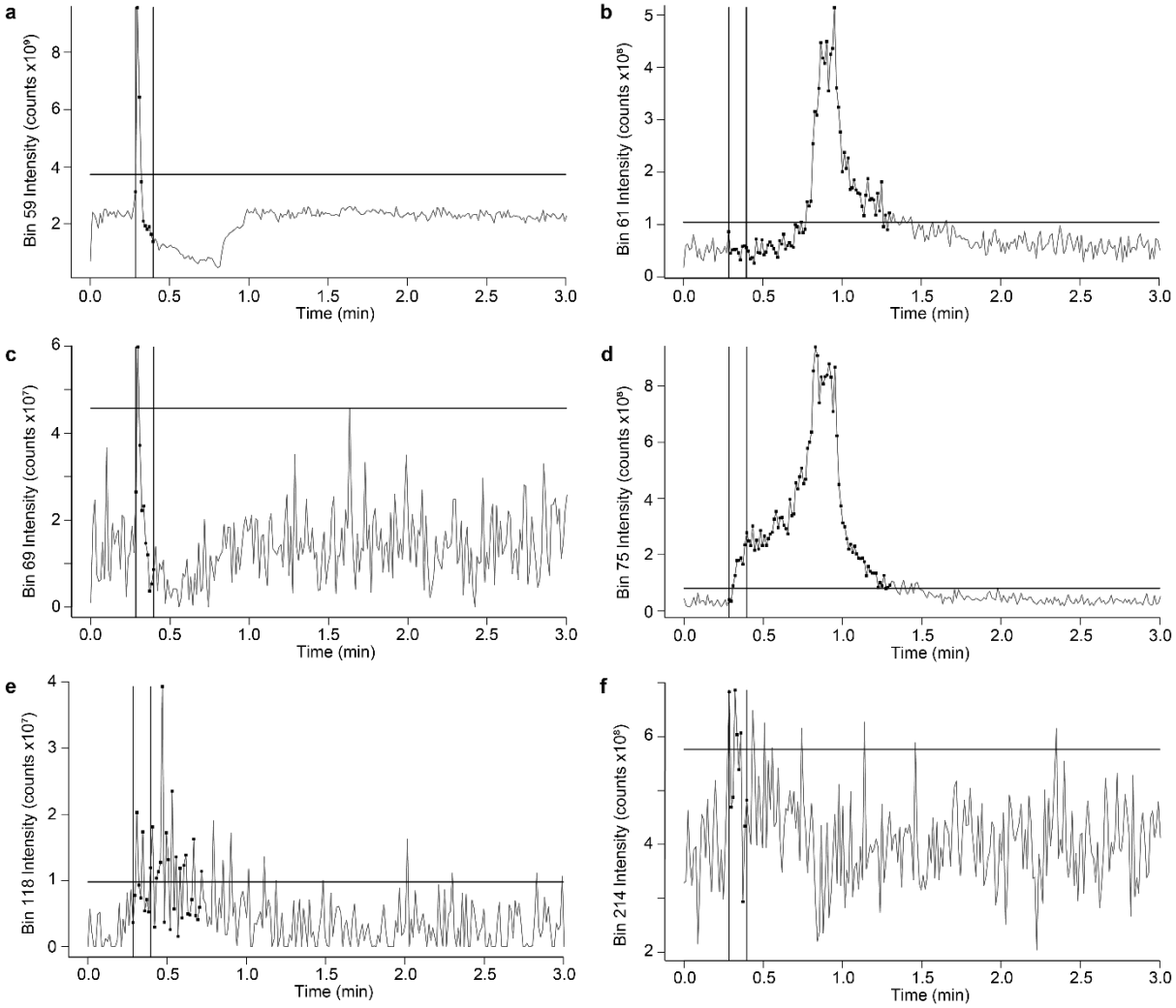

Figure 6. *The intensity over time of a single breath sample (grey line) for bins (a) 59, (b) 61, (c) 69, (d) 75, (e) 118, and (f) 214. The vertical black lines indicate the time of the first and tenth scans of breath, the horizontal line is the threshold (the mean plus three times the standard deviation of the first ten scans of ambient air), and the black dots indicate points that were selected by (4) Stable selection method. Therefore, data between the vertical lines was included in (1) Sum of the first 10 scans, data above the horizontal line was included in (3) Sum of all above the threshold, data between the vertical lines and above the horizontal line was included in (2) Sum of data points above the threshold in the first 10 scans, and all data after the first vertical line was included in (5) No selection.*

Isoprene usually gives a peak at m/z 69 (King et al., 2009), and this bin was included in every exercise versus resting model, and resulted in the best single-marker model. It has been shown previously that exercise affects isoprene levels, with a spike at the onset of exercise, followed by a decrease to below normal resting levels after around 10-15 min of continuous exercise (King et al., 2009; Schwoebel et al., 2011). The breath samples in this study were taken 12-27 min after the onset of exercise; however, the intensity increased every three minutes, differing from previous studies that were at a single, low intensity. However, similar results were found, with lower intensity measures of bin 69 during exercise compared to rest. It has been suggested that short-term changes in breath isoprene concentration are not due to differences in endogenous production but altered gas exchange conditions, which occur during exercise (King et al., 2009).

Despite these clear changes in bin 69, the addition of other potential biomarkers significantly improved the model. However, these additional ions may not necessarily be biologically relevant. Bins 91 and 98 were good single-

markers, with a Youden's index of ≥0.5 in both the training and validation sets.

Surprisingly, bin 59 (acetone) resulted in a fairly poor single-marker model, despite previously being shown to increase during continuous exercise (King et al., 2009; Senthilmohan et al., 2000). This may be due to the type of exercise performed; in this study, participants should have been at maximal oxygen uptake when the breath sample was collected, whereas in previous studies lower intensity continuous exercise was performed. Ohkuwa, Funada, and Tsuda (2011) found that acetone concentration in expired air increased with higher exercise intensity; however, their highest intensity (162 W) was much lower than the intensities performed in this study (405-585 W). As acetone is a product of lipolysis, the suppression of fat metabolism during near-maximal exercise (Achten, Gleeson, & Jeukendrup, 2002), may have prevented breath acetone levels from increasing in this study.

The methods in this paper do not allow for the identification of molecules. However, as the potential biomarkers selected by these methods correspond to previously identified biomarkers, this indicates that the methods are successfully detecting biologically relevant molecules.

During exercise, breathing patterns change dramatically, with quicker, deeper breaths. As the datapoint selection methods still managed to select data which could distinguish from resting breath, this suggests that differences in breathing depth and frequency do not diminish the performance of the method, particularly for (D) Peaks. However, none of the participants in this study had impaired respiratory patterns, so further investigation is required to determine the effects of different breathing patterns on the performance of the datapoint selection methods.

4.1.2 Baseline versus 120 min post-ingestion of peppermint oil models

(E) Mean 59 resulted in the worst prediction models for peppermint ingestion, with the lowest Youden's index, indicating that the few data points selected were not enough to obtain a representative sample for all the important metabolites. All the other datapoint selection models had a Youden's index of 1, indicating perfect separation of baseline and post-ingestion breath samples. The (F) No preselection model had a fairly high Youden's index of 0.962 and only marginally the highest residual deviance of all the peppermint models; nevertheless, the model still performed the second-worst. As in the exercise models, this suggests that selection methods A-D were successfully removing noise, although in this case there was probably initially less noise to remove.

Previously, targeted analysis of the peppermint breath dataset looked for and identified peaks at m/z 81, 137 and 155 (Heaney et al., 2016). These correspond to known ions for monoterpene fragmentation (m/z 81), oxygenated terpene fragmentation (m/z 137) and, and methone and eucalyptol (m/z 155 - where both are major components of peppermint oil). In this analysis, bins 138 and 82 were included in the prediction model and resulted in the best single-marker models. However, the data were grouped into 1 m/z bins, from –0.5 to +0.5 of each m/z, and examination of the original spectra revealed these peaks at m/z 137.6 and a combination of peaks at m/z 81.4 and 82.5. Therefore, a different binning method could have labelled these as m/z 81 and 137. Additionally, there were high correlations between bins 81 and 82 (r = 0.99) and bins 137 and 138 (r = 0.93), thus these results support the use of the protocols in this paper. Bin 152 also resulted in a good single-marker model, with further investigation required to assess whether this bin is biologically relevant.

Bin 155 (methone and eucalyptol) was not included in any of the models; however, the previous analysis of the data set showed that it peaked at 60 min post-ingestion in most participants. As the models were comparing baseline to 120 min post-ingestion of peppermint oil, the differences may be too small at that time to include bin 155. Additionally, lasso penalisation removes some variables that correlate highly with others. Bin 155 correlated with bin 138 (r = 0.77) which may be why it was excluded.

A variable selection step was not performed on this data, as an ambient air sample was not available. Whether this would improve the predictive models needs to be examined. Generally, participants are required to breathe while inhaling the filtered air for a few minutes prior to sample collection. This means that many VOCs from the ambient air will be washed out, and those that remain may not be from the air in that room. Therefore, comparing to an ambient air sample may not be relevant.

4.1.3 Continuous samples reliability

The reliability of the baseline continuous samples was improved by the datapoint selection methods. This indicates that data collected during inhalation added noise to the data, corroborating the results of the predictive models. The filtered air method generally gave lower RSD values than the ambient air method, as expected, because filtered air contains far fewer VOCs than ambient air and thus will reduce the noise. Both continuous methods reached the

generally accepted reliability of <20% (Kirwan et al., 2014).

Using the sum of randomly resampled data points will have improved the reliability. This is philosophically similar to standard methods for other types of MS data, in which intensities are obtained from samples over an extended time. Statistically, summing random time samples reduces noise, whether that is from the ambient air, random noise or instrument drift. The original datasets contain many zeros, likely from a lower measurement threshold, thus the sum of the random resamples has the additional benefit of reducing the likelihood of a zero measure. Furthermore, as each dataset can be resampled multiple times, this method can be used to increase the sample size to give multiple samples from across analytical variation within a single run, and thus reduce the chance of overfitting the model.

The effect of a shorter sampling time needs to be investigated. The (D) Peaks method selects few data points and this may result in too few to adequately sample across the noise. In this case, (B) Local regression may be more appropriate. The scan rate should also be considered before applying a selection method. A higher scan rate would produce more points during the exhalation that the (D) Peaks method would not select, but the other methods would.

This study is aiming to find a data processing method which produces the most reliable data and can detect differences between conditions. The methods have been considered with a mind to selecting the alveolar phase, with the graphs in Figure 2 and the model results suggesting that (D) Peaks may be achieving this. However, selection of breath phase was not the purpose of the study, and we did not have access to ground truth data to confirm it. The selection methods produced more accurate models than no preselection and improved reliability, thus achieving the aim. The methods can be applied to any real-time mass spectrometry analysis, although which methods are optimal may depend on scan rate and sampling time. The statistical approach used here for evaluating the methods, based on predictive function and reliability, is quite general.

4.2 Single breath study

The models to distinguish between breath and ambient air all resulted in perfect separation, demonstrating that clear differences between breath and ambient air can be detected using this sampling method.

The best reliability in the single breath samples was achieved by using (5) No selection; however, this included data from the end of the sampling period which would largely consist of ambient air. This was demonstrated by the far fewer potential biomarkers the method was able to detect compared to the datapoint selection methods. Therefore, this method would not be a method of choice for large range of applications. (1) Sum of the first 10 gave better reliability than the other selection methods. However, as shown in Figure 6, many of the potential breath biomarkers reached their maximum after the first 10 data points, so that method will not give an accurate representation of how much of each marker there is in each breath. Additionally, some barely increased compared to ambient air until after the first 10 data points, and therefore will not be selected by the algorithm as potential breath biomarkers, and thus may be excluded from an analysis where they may be of biological relevance.

(2) Sum of points above the threshold in the first 10 points and (3) Sum of all above the threshold both assume that any data point above the threshold is due to the breath. As any points below the threshold are discarded, it is therefore also assumed that any points below the threshold are not breath, which may not be true, particularly in cases where it is above the mean of the air measure. For example, if an intensity is just above the threshold, this will be included in the final sum, but if it is just below, it will not. Therefore, small differences in individual points can each change the final sum by the value of the threshold, for (2) Sum of points above the threshold in the first 10 points and (3) Sum of all above the threshold. This issue was likely one reason why the reliability was so poor for these methods.

Another potential issue for the threshold methods (2 and 3) is when all data points for a bin within the ambient air measure are zero. As the zeros are removed, this results in a null threshold and the ion is considered not present in that sample, regardless of the intensity measured in the breath. Obviously, this could be completely inaccurate, but fairly likely for molecules which are scarce in ambient air and common in breath. Simply setting the threshold as zero would be a simple fix for this issue; however, any small amount of noise would suggest that the ion is present in breath.

(4) Stable selection method aims to fix the issues within the other methods. First, by not limiting the measurement to the first 10 data points, ions that appear or reach maximum later are measured more accurately. Second, it does not discard measurements based only on whether they are below the threshold, but on whether they are consistently below the threshold. A small amount of noise could put a measure below the threshold, but it is unlikely that this will occur for more than five points in a row, so potential biomarkers with low intensities that hover just around the threshold can still be identified. Third, if all data points for a bin within the ambient air measure are zero, the threshold

is the standard deviation of the whole sample rather than zero. This reduces the chance of a non-biomarker channel being selected due to random noise above the lower limit of detection (zero). This method gave better reliability than the threshold methods, suggesting that discarding any measure below the threshold could increase noise in the measurement.

It was expected that all the breath would leave the tube in around 7 s (10 scans). However, Figure 6 demonstrates that some show an increase in intensity much later. This may be due to how volatile the compounds are, and some may attach to the tube or sample line material. This is effectively providing a form of separation. However, it makes selecting the correct data points more difficult, as it is unclear when each bin will have completely left the tube. (4) Stable selection method should be able to detect potential markers that are measured later, while avoiding selecting those that are simply noise. However, this method has only been tested on one participant under the same conditions, further investigation of the method, including creating multivariate models to distinguish between conditions, is required to determine whether further improvements are necessary.

This study only examined one participant. It is possible, but unlikely, that the best breath selection method will be different for different participants under the same conditions. However, most of the issues discussed in this paper are likely to remain, regardless of any differences in breath composition. Additionally, the samples were taken across several days, so there would be some day-to-day variation in the samples.

Twelve potential breath biomarkers were identified by all the methods applied to the single breath data (bins 59, 60, 61, 75, 76, 77, 79, 80, 89, 90, 95 and 231), and an additional 101 identified by at least one of the methods. Further work should be done to identify whether any of the detected channels are from contaminants, for example from the tube material. Of the 20 potential breath biomarkers identified by the (D) Peaks method in the exercise study resting data, 14 of these were also identified by the (4) Stable selection method in the single breath data (bins 37, 41, 59, 60, 63, 64, 68, 69, 70, 75, 76, 89, 91, and 232). These include acetone (59), dimethyl sulphide (63), isoprene (59) and protonated water cluster (37), which have been previously identified as components of breath (Moser et al., 2005; Schwoebel et al., 2011).

Currently, these methods are semi-quantitative in that they allow for both qualitative analysis: is a bin present, and quantitative analysis of the intensity measurement. That is, the results are quantitative in a statistical sense, though not in a chemical sense; the measured intensities have some meaning in terms of prediction of participant state, which is the aim here. However, the relationship to the quantity of a VOC present in the sample is not known.

5. Conclusion

The method of selecting exhaled breath from CMS data can affect the reliability of the measurement and the ability to distinguish between breath samples taken under different conditions.

The results here suggest that for continuous samples taken while the participant is inhaling ambient air, choosing the peak of each breath will provide the best results, and for samples taken while the participant is inhaling filtered air, any of the datapoint selection methods presented here based on total intensity will provide good results. Further research is necessary to ensure the breath selection methods are not affected by differences in respiratory patterns, and to optimise the variable selection criteria.

For the single breath samples, the new method proposed here will select potential markers, regardless of when they are measured, with reasonable reliability. Further investigation is necessary to confirm the biological relevance of the potential biomarkers and the performance of the method on experimental data.

The application of appropriate data processing methods can improve the quality of the data and results obtained from CMS. The methods presented will enable untargeted analysis of breath VOCs using CMS to be performed.

Acknowledgments

The authors would like to thank Dr Liam Heaney and colleagues, for allowing the data from the study published in Journal of *Bioanalysis*, (Heaney et al., 2016) to be used in this paper.

References

Achten, J., Gleeson, M., & Jeukendrup, A. E. (2002). Determination of the exercise intensity that elicits maximal fat oxidation Shifts in energy substrate mobilization and utilization. *Medicine & Science in Sports & Exercise*, *34*(1), 92–97. Retrieved from https://pdfs.semanticscholar.org/63a0/895c7f543b0a27422fd8ea3fc294e4afd475.pdf

Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 13: Receiver operating characteristics curves. *Critical Care*, *8*(6), 508–512. https://doi.org/10.1186/cc3000

Bouza, M., Gonzalez-Soto, J., Pereiro, R., De Vicente, J. C., & Sanz-Medel, A. (2017). Exhaled breath and oral cavity VOCs as potential biomarkers in oral

cancer patients. *J. Breath Res*, *11*. Retrieved from https://doi.org/10.1088/1752-7163/aa5e76

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22. Retrieved from http://www.jstatsoft.org/v33/i01/.%0D

Heaney, L. M., Ruszkiewicz, D. M., Arthur, K. L., Hadjithekli, A., Aldcroft, C., Lindley, M. R., … Reynolds, J. C. (2016). Real-time monitoring of exhaled volatiles using atmospheric pressure chemical ionization on a compact mass spectrometer. *Bioanalysis*, *8*(13). https://doi.org/10.4155/bio

King, J., Kupferthaler, A., Unterkofler, K., Koc, H., Teschl, S., Teschl, G., … Amann, A. (2009). Isoprene and acetone concentration profiles during exercise on an ergometer. *Journal of Breath Research*, *3*(2). https://doi.org/10.1088/1752-7155/3/2/027006

Kirwan, J. A., Weber, R. J. M., Broadhurst, D. I., & Viant, M. R. (2014). Direct infusion mass spectrometry metabolomics dataset: A benchmark for data processing and quality control. *Scientific Data*, *1*, 1–13. https://doi.org/10.1038/sdata.2014.12

Liland, K. H. (2011). Multivariate methods in metabolomics - from pre-processing to dimension reduction and statistical analysis. *TrAC - Trends in Analytical Chemistry*. https://doi.org/10.1016/j.trac.2011.02.007

McNeish, D. M. (2015). Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences. *Multivariate Behavioral Research*, *50*(5), 471–484. https://doi.org/10.1080/00273171.2015.1036965

Moser, B., Bodrogi, F., Eibl, G., Lechner, M., Rieder, J., & Lirk, P. (2005). Mass spectrometric profile of exhaled breath - Field study by PTR-MS. *Respiratory Physiology and Neurobiology*, *145*(2–3), 295–300. https://doi.org/10.1016/j.resp.2004.02.002

Ohkuwa, T., Funada, T., & Tsuda, T. (2011). Acetone Response with Exercise Intensity, (2006).

Perez-Hurtado, P., Palmer, E., Owen, T., Aldcroft, C., Allen, M. H., Jones, J., … Reynolds, J. C. (2017). Direct analysis of volatile organic compounds in foods by headspace extraction atmospheric pressure chemical ionisation mass spectrometry. *Rapid Communications in Mass Spectrometry*, *31*(22), 1947–1956. https://doi.org/10.1002/rcm.7975

Pizzini, A., Filipiak, W., Wille, J., Ager, C., Wiesenhofer, H., Kubinec, R., … Bellmann-Weiler, R. (2018). Analysis of volatile organic compounds in the breath of patients with stable or acute exacerbation of chronic obstructive pulmonary disease. *Journal of Breath Research*, *12*(3), 36002. https://doi.org/10.1088/1752-7163/aaa4c5

Poli, D., Goldoni, M., Corradi, M., Acampa, O., Carbognani, P., Internullo, E., … Mutti, A. (2010). Determination of aldehydes in exhaled breath of patients with lung cancer by means of on-fiber-derivatisation SPME-GC/MS. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*. https://doi.org/10.1016/j.jchromb.2010.01.022

R Core Team. (2013). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Schwoebel, H., Schubert, R., Sklorz, M., Schubert, J. K., & Miekisch, W. (2011). Phase-resolved real-time breath analysis during exercise by means of smart processing of PTR-MS data, 2079–2091. https://doi.org/10.1007/s00216-011-5173-2

Senthilmohan, S. T., Milligan, D. B., McEwan, M. J., Freeman, C. G., & Wilson, P. F. (2000). Quantitative analysis of trace gases of breath during exercise using the new SIFT–MS technique. *Redox Report*, *5*(2–3), 151–153. https://doi.org/10.1179/135100000101535537

Trefz, P., Schmidt, M., Oertel, P., Obermeier, J., Brock, B., Kamysek, S., … Miekisch, W. (2013). Continuous real time breath gas monitoring in the clinical environment by proton-transfer-reaction-time-of-flight-mass spectrometry. *Analytical Chemistry*, *85*(21), 10321–10329. https://doi.org/10.1021/ac402298v

Turner, C. (2016). Techniques and issues in breath and clinical sample headspace analysis for disease diagnosis. *Bioanalysis*. https://doi.org/10.4155/bio.16.22

Van Berkel, J. J. B. N., Dallinga, J. W., Möller, G. M., Godschalk, R. W. L., Moonen, E. J., Wouters, E. F. M., & Van Schooten, F. J. (2010). A profile of volatile organic compounds in breath discriminates COPD patients from controls. *Respiratory Medicine*. https://doi.org/10.1016/j.rmed.2009.10.018

Wehinger, A., Schmid, A., Mechtcheriakov, S., Ledochowski, M., Grabmer, C., Gastl, G. A., & Amann, A. (2007). Lung cancer detection by proton transfer reaction mass-spectrometric analysis of human breath gas. *International Journal of Mass Spectrometry*. https://doi.org/10.1016/j.ijms.2007.05.012

Wenig, P., & Odermatt, J. (2010). OpenChrom: A cross-platform open source software for the mass spectrometric analysis of chromatographic data. *BMC Bioinformatics*, *11*(1), 405. https://doi.org/10.1186/1471-2105-11-405