Statistical analysis of MCC-IMS data for two group comparisons—an exemplary study on two devices

S Horsch, J I Baumbach and J Rahnenführer

1 Department of Statistics, TU Dortmund, D-44221, Dortmund, Germany
2 B. Braun Melsungen AG Center of Competence Breath Analysis, D-44227, Dortmund, Germany
3 Former address: Faculty of Applied Chemistry, Reutlingen University, D-72762, Reutlingen, Germany

E-mail: rahnenfuehrer@statistik.tu-dortmund.de

Keywords: MCC-IMS, confounders, scaling, classification, automatic peak detection

Abstract

The Multi-capillary-column-Ion-mobility-spectrometry (MCC-IMS) technology for measuring breath gas can be used for distinguishing between healthy and diseased subjects or between different types of diseases. The statistical methods for classifying the corresponding breath samples typically neglects potential confounding clinical and technical variables, reducing both accuracy and generalizability of the results. Especially measuring samples on different technical devices can heavily influence the results. We conducted a controlled breath gas study including 49 healthy volunteers to evaluate the effect of the variables sex, smoking habits and technical device. Every person was measured twice, once before and once after consuming a glass of orange juice. The two measurements were obtained on two different devices. The evaluation of the MCC-IMS data regarding metabolite detection was performed once using the software VisualNow, which requires manual interaction, and once using the fully automated algorithm SGLTR-DBSCAN. We present statistical solutions, peak alignment and scaling, to adjust for the different devices. For the other potential confounders sex and smoking, in our study no significant influence was identified.

1. Introduction

Breath gas analysis has shown great potential for diagnosing diseases [1, 2]. However, often the potential of the data to discriminate patient groups is not fully exploited, since only the metabolites identified in the breath are analyzed and possible confounders such as clinical variables are ignored. This is surprising, since several studies have discovered relationships between the composition of human breath and such confounding variables, e.g. sex or smoking status [3, 4]. This can be a serious problem, especially when the confounding variable is related to the presence of the disease and the study design is not carefully chosen [5].

1.1. Confounders

An insightful example is given by lung diseases, such as lung cancer or COPD, in case control studies. Here, the risk is high that diseased probands are smokers, whereas controls are younger and more likely to be non-smokers. It is known that smokers can be detected through different concentrations of specific volatile organic compounds (VOCs) in their breath air [3]. Therefore, in a study comparing smoking diseased probands with healthy non-smokers, it would be sufficient for a statistical classification algorithm to differentiate between smokers and non-smokers in order to achieve perfect accuracy. When applying the corresponding statistical model to data from the real world, false negatives would arise for non-smoking diseased subjects and false positives for healthy smokers.

Another example for a possible confounder is the sex of a proband. It has been shown that there are differences in the breath gas composition between men and women and also that men and women react differently to the exposition of certain substances [6, 7]. In our study, we focus on these two variables as potential confounders and do not consider hidden confounders which are not directly observed. In [8] an open source software was introduced that can identify similar groups of variables or subjects by applying cluster algorithms and that can help to identify hidden confounders.
1.2. Technology

Multi-capillary-column-Ion-mobility-spectrometry (MCC-IMS), as reviewed in [9], is one of several technologies for analyzing breath gas. It operates in two consecutive separation steps. In the first step, the molecules are separated by passing the multi-capillary-column (MCC). The time a molecule needs to pass the column is called retention time (RT). Afterwards, arriving molecules are transferred into the ion mobility spectrometer (IMS), where they get ionized and lead through a drift tube. The time needed to pass the drift tube is called drift time, which is transformed into a quantity called IRM inverse reduced mobility (IRM). The combinations of the two measured times are specific for certain VOCs, that can be identified later, e.g. by an experiment or database information. The amount of a certain metabolite is measured at the end of the drift tube, where the ions meet a faraday plate. The resulting electrical potential is considered the intensity of the respective compound. The advantages of MCC-IMS are the ability to cope well with the humidity of human breath, the fast evaluation (about 10–15 min for the analysis of a breath sample) and increasing miniaturization and mobility of the device, allowing its utilization e.g. in hospitals, where limited space for equipment is always a critical issue in establishing new instruments.

1.3. A controlled study

We conducted a controlled study including 49 healthy subjects in order to analyze the potential of statistical methods that deal with confounders. In the study we examined, if sex and smoking status can be distinguished using MCC-IMS. Furthermore, there is often just one device used in MCC-IMS studies. Although this eliminates a source for variability in the data of these studies, it is an unrealistic scenario for broad applications, where many devices are used. It remains unclear, if results from one device can directly be transferred to another. Therefore, we used two devices, A and B, for the analysis. In order to create a scenario with different groups within our study population, each subject was measured twice, once before and once after drinking a glass of orange juice. Drinking before the measurement can itself be a typical confounding variable. Here, the two measurements are used to generate two groups, mimicking a study with healthy and with diseased subjects.

In order to evaluate if differences exist regarding the devices, the sex and the smoking status of the subjects, several analysis steps have to be undertaken. After collecting the MCC-IMS raw data it is necessary to define the compounds in the air by examining the heat maps, displaying the IRM and the RT on the x-axis and and y-axis, respectively, and the signal intensity by color. Compounds can visually be spotted in the image by peak areas of higher intensities separating from the background. This assessment can be carried out semi-manually by experts, e.g. using the software VisualNow as a gold standard. Recently, fully automated algorithms were successfully developed and applied, some of them achieving at least equal results in a subsequent classification task [10]. The suggested combination of algorithms, called SGLTR-DBSCAN, detects so-called single peaks in each measurement separately and afterwards clusters the peaks among several measurements, defining a common list of so-called consensus peaks. The manual approach does not divide this process into two steps but delivers the consensus peak list directly. We descriptively compare the results of the manual and automatic peak detection and use the results of the classification task as an indicator for successful peak detection.

Since it is known that peak positions of the same compound might differ on the two devices [11], we use results from a database, where synthetic combinations of known substances were measured on both devices, five of them intersecting. In order to transform the peak positions from one device into positions matching the other device, we suggest a linear regression of both, the RT and the IRM, of the five peak positions from the synthetic gas. The peak positions of all single peaks from our device B are thus aligned to the single peak positions of our reference device A. Afterwards, the peak clustering step is applied. Different scaling methods are proposed to further reduce the effect of the device.

1.4. Analysis for group comparisons

We first compare the means of the concentrations of all detected metabolites for the groups with/without juice, device A/B, male/female and (ex-) smoker/non-smoker with (nonparametric) univariate (paired) Wilcoxon tests and, depending on the testing hypothesis, (parametric) Welch- or paired t-tests. The respective p-values are adjusted by the method of Bonferroni–Holm, controlling the global significance level $\alpha = 0.05$, separately for each two-group comparison. In a multivariate approach we then statistically classify two groups (with/without juice) using the statistical learning algorithm Random Forest (RF), which has proven to be a suitable algorithm for this task, outperforming popular alternatives like Support Vector Machines, k-Nearest-Neighbor-Algorithms or simple Classification Trees [10, 12]. Classification is applied in a 50 times repeated stratified 10 fold cross-validation setting. As performance measure, the Area under the Receiver Operating Characteristic Curve (AUC) is used.

Further, in all analyses we compare the results obtained with manual peak annotation with the results obtained with automated peak detection. For demonstration purposes, we report the impact on the classification results, when the data from one device are used for training and the data from the other device are used for testing, both with and without alignment and scaling.
1.5. Structure of the paper
In the next section, the design of the controlled study and the metadata of the study are presented. In section 3, the MCC-IMS technology is described in detail. In section 4, all methods for preprocessing the data including peak picking, peak alignment, peak clustering, and scaling, as well as the methods for univariate comparisons of compound concentrations and multivariate methods for classification are presented. Section 5 contains the results. First, the alignment of peak positions after preprocessing and the impact of the scaling methods are discussed. Then, univariate tests for metabolites are presented, on the one hand to identify discriminating metabolites between the groups with and without juice and to identify metabolites that indicate differences with respect to confounding factors. Finally, the impact of methods for peak identification and scaling on classification results are investigated.

2. Controlled study and metadata
The data were collected in May 2017 at B&S Analytik in Dortmund, Germany. The participants were mostly staff from the Technical University of Dortmund and from the laboratory of B&S Analytik. There were no exclusion criteria except for pregnancy (due to laboratory regulations) and allergies to citrus fruits, since each participant consumed a glass of orange juice in the course of the study. Participants were not pre-selected according to sex or smoking status. Out of the 49 volunteers, 26 were male and 23 female. The age ranged from 23 to 49 years (median 29, MAD 2.97). 40 volunteers never smoked regularly, 5 had smoked regularly in the past but quit and 4 were still smoking. Each participant answered a questionnaire about their sex, age, smoking habits, food consumption, and drinks in the last 12 h before the experiment and signed an agreement to participate in the study. Each participant’s breath was measured on two different MCC-IMS devices, first on one device before drinking a glass of orange juice and then on the other device after drinking. This design was chosen in order to use paired samples for the target variable juice and at the same time observe the effect of the device. The starting device was varied such that not all measurements with juice were measured on the same device which would make it impossible to distinguish the effect of the device from the effect of the juice. The starting device was determined by stratified randomization with regard to the sex of the participant. As a result, 11 women and 13 men were measured first by device A and 12 women and 13 men were measured first by device B. Along with each analysis of breath gas comes a subsequent measurement of the room air and a measurement taken during the cleaning flush at the end of the single measurement. This data is not investigated here, except that the peak detection process is based on all available measurements.

3. MCC-IMS technology
Two MCC-IMS devices from B&S Analytik were used for data acquisition. Device A uses software version VOCan—v2.7, device B VOCan—v3.4.1. Both devices work at ambient temperature and pressure. They use an OV5-MCC (multi-capillary-column), operating at 40 °C. The breath sample (10 ml) enters the MCC, where a pre-separation of contained molecules is achieved. The time a molecule takes to pass the MCC is measured and called RT. Afterwards, the molecules are lead into the IMS, where they get ionized and enter the drift chamber, controlled by a shutter. The grid opening time of the shutter is 300 μs. The drift length is 120 mm. The electrical potential caused by the drift rings along the drift tube leads the ions along the tube. The time the molecules need to pass the drift tube is called DT (drift time), which is then transformed into the IRM. The carrier gas flow rate is 150 ml min$^{-1}$, the sample and the drift gas flow rate 100 ml min$^{-1}$. The ions finally meet the faraday plate, resulting in an electrical potential (signal intensity) representing the amount of the respective molecule. The three quantities RT, IRM and signal intensity can be visualized in a heat map, thus displaying an entire measurement. An example for a raw measurement is displayed in figure 1. Increasing signal intensity is displayed by a color gradient from light blue to dark blue, red and yellow.

4. Methods
The statistical analyses and the alignment of the peak positions were performed with the software R [13]. The automatic peak picking (SGLTR) and the peak clustering (DBSCAN) were computed with software provided by the project group suggesting these algorithms for MCC-IMS peak detection [14].

4.1. Automated peak picking
The detection of peaks in a single measurement, which we call peak picking, is accomplished by the algorithm Savitzky–Golay Laplace-operator filter thresholding regions (SGLTR) [14]. It identifies peaks as regions with high curvature. The Laplace-operator, the sum of the second derivatives in drift and RT, is approximated by a weighted Savitzky–Golay filter that is based on a local polynomial regression. Pixels in the heat map with high curvature that exceed a certain threshold are identified as peak candidates. Connected peak candidates are merged to a peak region. The final peak position and its intensity are set to the corresponding values of the pixel with the highest intensity in the peak region. We call peaks resulting from this step single peaks.
4.2. Alignment of peak positions

It is known that the position of a peak induced by a certain metabolite varies among different MCC-IMS devices, also dependent from their settings, like temperature or pressure or even the time of the measurement \[11\]. Therefore, a reference composition, containing intersecting analytes, has been evaluated on both devices. We apply a linear regression for both dimensions, RT and DT, regressing the positions of the reference compounds from one device on the positions from the other device. Afterwards, the regression model is used to transform the single peak positions obtained from one device to the scale of the other device.

4.3. Automated peak clustering

Peak positions from several measurements always vary due to technical imprecision and biologic variability, even when they originate from the same metabolite in the breath and when they are measured by the same device. An algorithm is required in order to determine, which peaks from several measurements represent the same chemical compound and thus must be merged to one so-called consensus peak. This step, which we call peak clustering, is accomplished by Density-Based Spatial Clustering of Applications with Noise (DBSCAN). It finds clusters in a sequential procedure. First, peaks that have at least $p_{\text{min}}$ (here 10) peaks in a certain radius around its position, are labelled as potential consensus peaks. All peaks in this radius are then merged to a cluster. Then, it is iteratively checked for all peaks in the cluster, if additional peaks lie within their respective radiuses. Those peaks are also added to the cluster, and so forth, until no peaks can be added to an existing cluster. The final peak position is set to the centroid of the cluster.

4.4. Scaling methods

In order to match the signal intensities of a metabolite measured on two devices, scaling methods are applied to the values of both devices separately. Let $x_1, \ldots, x_n$ be the values of one metabolite for device A and $y_1, \ldots, y_n$ the values for the same metabolite on device B. Then the scaled observations for both devices are calculated as follows, where $\bar{x} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i$ and $\bar{y} = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i$ are the arithmetic means for both devices, $\hat{\sigma}_x = \sqrt{\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2}$ and $\hat{\sigma}_y = \sqrt{\frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_i - \bar{y})^2}$ are the standard deviations for both devices and $\hat{x}^{(a=0)}$, $y^{(a=0)}$, $\hat{\sigma}_x^{(a=0)}$, $\hat{\sigma}_y^{(a=0)}$ the same quantities calculated after removing all values that are zero from the dataset.

(i) Mean scaling (scaling the data to mean zero, separately for both devices):

$$x_i^{(\text{m0})} = x_i - \bar{x}, \quad y_i^{(\text{m0})} = y_i - \bar{y}$$

(ii) Variance scaling (scaling to standard deviation one, separately for both devices):

$$x_i^{(\text{v1})} = \frac{x_i}{\hat{\sigma}_x}, \quad y_i^{(\text{v1})} = \frac{y_i}{\hat{\sigma}_y}$$

(iii) Normalization (scaling to mean zero and standard deviation one separately for both devices):

$$x_i^{(\text{n0v1})} = \frac{x_i - \bar{x}}{\hat{\sigma}_x}, \quad y_i^{(\text{n0v1})} = \frac{y_i - \bar{y}}{\hat{\sigma}_y}$$

In the results section, this type of normalization is called scale.

(iv) Normalization with zeros shifted equally (like (iii) but only for the values that are not zero. The zeros

![Figure 1. Example for a raw MCC-IMS measurement. The x-axis denotes the transformation of the drift time, the y-axis denotes the retention time and the colors the respective intensity. Increasing intensity is displayed by a color gradient from light blue to dark blue, red and yellow. Neighboring pixels with similar intensities form visually recognizable peaks, representing a common originating metabolite. The very large intensities at IRM = 0.5 belong to the reactant ion peak, an artifact that is removed during the peak analysis.](image-url)
from both devices are shifted by the same value, such
that they are certainly lower than any other scaled
value that was not zero before):

\[
x_i^{(\text{min}^{+})} = \begin{cases} 
  \frac{x_i - x_{i}^{0}}{\delta_{x}^{0}}, & \text{for } x_i 
eq 0, \quad \delta_{x}^{0} 
eq 0 \\
  \frac{x_i - x_{i}^{0}}{\delta_{x}^{0}} + \max\{|x_i - \bar{x}_{i}^{0}|, |\bar{x}_{i}^{0} - \bar{x}_{i}^{0}|\}, & \text{for } x_i = 0, \quad \delta_{x}^{0} = 0 \\
  \max\{|x_i - \bar{x}_{i}^{0}|, |\bar{x}_{i}^{0} - \bar{x}_{i}^{0}|\}, & \text{for } x_i = 0, \quad \delta_{x}^{0} = 0 \text{ and } \delta_{y}^{0} = 0 \\
  \max\{|x_i - \bar{x}_{i}^{0}|, |\bar{x}_{i}^{0} - \bar{x}_{i}^{0}|\}, & \text{for } x_i = 0, \quad \delta_{x}^{0} = \delta_{y}^{0} = 0 \\
\end{cases}
\]

\( y_i^{(\text{min}^{+})} \) is defined analogously.

In the results section, this type of normalization is
called scale+.

(v) Robust normalization with zeros shifted equally
(like (iv) but robust for outliers): \( x_i^{(\text{min}^{+})} \), \( y_i^{(\text{min}^{+})} \)
are analogously defined to (iv), but replacing
means by medians and standard deviations by
median absolute deviations (MAD).

4.5. Tests for mean differences

To test which features (metabolites) can univariately
discriminate between two groups, we apply both a
parametric test (the Welch-test, a version of the \( t \)-test,
where the standard deviations in both groups are not
assumed to be equal) and a non-parametric test
(Wilcoxon test). To test differences between the two
groups ‘with juice’ and ‘without juice’ or between
the groups ‘device A’ and ‘device B’ there are two
measurements available per subject, one in each
condition. In these cases, paired \( t \)-tests and Wilcoxon
tests are performed. In order to control the type I
error, we adjust the \( p \)-values for each group compar-
ison according to the method of Bonferroni–Holm.
We use \( \alpha = 0.05 \) as significance level.

4.6. Classification methods

For statistical classification we use a RF [15], which
proved to be a favorable choice in MCC-IMS two-class
tasks [10, 12]. A RF builds a decision model by
averaging many (here 1000) Classification Trees. Each
tree iteratively divides the observations by binary splits
of single variables into so-called nodes such that the
values of the variables in both groups are as separated
as possible by maximizing the Gini-Index in the new
nodes. Each resulting node is further split by another
binary split until the observations are as perfectly as
possible divided into terminal nodes. For each split
the best variable (metabolite) out of \( \sqrt{p} \) randomly selected
variables is chosen, where \( p \) denotes the total number
of variables, here identified metabolites (peaks), in the
respective dataset. When a new observation is classi-
fied, it is passed through all 1000 Classification Trees.
The percentage of trees voting for a certain class is the
estimated probability for the respective class. The
observation is then assigned to the class with prob-
ability greater than 0.5. The importance value (mean
decrease of the Gini-Index), a measure for decreased
impurity of nodes created by a split, allows to quantify
the relevance of single metabolites for a successful
classification. In order to achieve unbiased estimates and
observe the variability of the results, a 50 times repeated
10 fold stratified cross-validation is performed.

As performance measures, we focus on the AUC
and the mean classification error (MMCE), addition-
ally reporting the false positive rate (FPR), false nega-
tive rate (FNR), positive predictive value (PPV) and
negative predictive value (NPV).

5. Results

5.1. Impact of peak identification

The first pre-processing step of the raw MCC-IMS
measurements is peak identification. We define three
different peak identification methods (manual, auto-
matic and automatic with alignment) and compare
them on the data.

In total, 345 raw MCC-IMS-measurements are
available, including 98 breath gas measurements (two
per subject), 98 room air measurements (one per breath
gas measurement) and 149 flush measurements (one per
breath gas measurements plus some additional clean-
ing). For the peak identification step, all these
measurements were used, in order to perform this step as stable
as possible. Afterwards, the analysis and the evaluation
are restricted to the breath gas samples.

Peak identification with the semi-manual proce-
dure using the software VisualNow is denoted by manual
and leads to 124 peaks. Automatic peak
identification, combining a peak picking and a peak
clustering step, is denoted by automatic and yields 51
peaks. The third peak picking algorithm combines
automatic with an alignment of peak positions regard-
ing the two devices and is called automatic with align-
ment). This method yields 46 peaks. The peak
positions vary between the two devices called A and B.
Therefore, in automatic with alignment a peak align-
ment is performed, shifting the positions of the single
peaks, i.e. the peaks identified in one single measure-
ment without assigning them to a consensus peak,
from the device B to the (arbitrarily chosen) reference
device A. The resulting regression coefficients for our
data are shown in table 1. For the RT, the shift is more pronounced than for the IRM.

The effect of the alignment is visualized in figure 2. On the left, the locations of the consensus peaks from the peak detection automatic are shown, on the right the locations from automatic with alignment. The raw measurements are not displayed here, for a clearer presentation. For each peak position it is noted, in how many of the raw measurements this peak was found, i.e. it has a value >0, both for device A and for device B. Room air and flush measurements are still included here. Red color indicates more non-zero observations on device A, blue more on device B. Often, red and blue dots with a high color intensity are close to one another (see for example at IRM times of about 0.6), suggesting that the true metabolite underlying the two peaks is actually the same. After the alignment (right) many imbalances are successfully compensated.

The sheer number of found peaks is not informative about the quality of the peak detection. Therefore, the final peak positions for all three peak identification methods (manual, automatic and automatic with alignment) are displayed in figure 3. The background shows the in total 345 averaged raw measurements. For all columns, the 20%-quantile intensity is subtracted from the intensities (resulting in negative values that are then set to zero) in order to remove noise.

Apparently, the automatic peak identification misses some of the visually recognizable peaks (e.g. at IRM ≈ 0.95 and RT ≈ 600), whereas the manual peak identification indicates peaks at positions, where no peak can be seen in this data representation (e.g. three peaks with IRM about 0.98). These problems can result (for automatic peak detection) from peaks not being found (often enough) in the single measurements and (for manual peak detection) from the use of layers (additional prior information about peaks found in other experiments). Two small, close peaks in the unaligned dataset (IRM ≈ 0.9 − 0.95, RT ≈ 100) are not identified until the peak positions are aligned (at the same position in the bottom figure), demonstrating the superiority of the alignment step.

Principal Component Analysis (PCA) allows the visualization of an entire dataset as a scatter plot, by reducing the dimension of the dataset to so-called principal components. They are linear combinations of the original variables, which are constructed in a way that they explain as much variability in the data as possible. Figure 4 shows the first two principal components for each dataset version (only breath data) and the explained proportion of variance of the represented dataset. Red symbols represent measurements from device A, blue symbols from device B. Circles mark measurements before drinking a glass of orange juice and triangles afterwards. Obviously, the device has a strong impact on the data and accounts for a considerable amount of variance in the data, since the colored symbols are separated, irrespective of the data version. There is also a clear difference between the measurements with and without juice.

5.2. Impact of scaling
In order to reduce the effect of the device, we suggest different scaling methods, among them commonly used mean scaling, variance scaling and their combination. Due to the particular situation, that the application of the automatic peak detection results in many values of exactly zero, when the algorithm did not detect a peak at this position, we introduce two
scaling functions tailored specifically to this scenario. First, they scale all values that are not zero to mean zero and variance one or median zero and median absolute deviation (MAD) one, respectively. These scaling functions are applied to the data separately for each device. Second, for all values that are exactly zero, the
values from both devices are shifted identically, such that they are always lower than the smallest non-zero value. This can be achieved by subtracting the larger mean/median and dividing by the smaller standard deviation/MAD. In the manual version of the dataset, these two methods have negligible effect, because there are almost no zeros. During the manual peak detection, for each measurement, the highest signal intensity in the peak area is chosen as observed value, whereas the automatic algorithm always returns zero, when the algorithm does not detect a peak at this position in the respective measurement.

The effect of the different scaling methods can be observed in figure 5. On top, a metabolite from the manual dataset is shown (red: device A, blue: device B, circle: without juice, triangle: with juice). The x-axis denotes the indices of the 98 observations and the y-axis the signal intensity of the metabolite. On the original scale, observations with and without juice have different means for both devices (with juice yielding higher values), but the variance on device B is quite larger. It can easily be seen that adjusting only for the mean or the variance (top center and right panel) is not sufficient for matching the devices. Adjusting for both mean and variance (lower left panel), the distributions of both devices look much more alike and the measurements for with/without juice also match much better. The bottom of figure 5 shows a metabolite from the aligned automated dataset, illustrating the issue of many values that are exactly zero. Due to the different numbers of zeros on both devices, aligning for mean zero, variance one or both, the adjustment fails regarding interpretation. Although both devices concluded, that the metabolite was not present (i.e. zero) in some measurements, the values are different after scaling each device separately. When only the variance is considered, the non-zero values from both devices are shifted away from each other to even out the variance. In this case, the version treating the zeros from both devices together but separately from the non-zero values achieves more intuitive results.

In this dataset, we did not observe extreme outliers, such that there was no big difference in using the robust alternative, i.e. median and MAD instead of mean and standard deviation. When outliers occur, we suggest to use the robust alternative. Here, we limit our further investigation to the two scalings with mean zero and standard deviation one (scale) and the one with additionally treating the zeros separately (scale*).

The effect of these two scalings on the PCA are displayed in figure 6 for the data with alignment. For the scaling including the zeros (left, scale), the effect of the device is almost invisible (although differences can still be seen at a closer look for the observations without juice). This is the expected result, since the mean and the variance were explicitly set to equal values, respectively. For the altered scaling with zeros treated separately (right, scale*), the differences are still present in the PCA (explained variance by the second principal component is reduced from 15% to 10% in comparison to the unscaled data in figure 4). Excluding the zeros from the scaling consequentialy results in different means and standard deviations for the two devices. Preserving the logic, that zeros on both devices should have the same value after scaling, this scaling method means accepting, that there are actual differences between the devices.
5.3. Univariate analyses for the identification of differential metabolites

In order to assess if there are differential metabolites regarding our group variable (with/without juice, replacing disease states) or the possible confounders (device, sex, smoking status), univariate tests for differences in location are performed. For the variables juice and device, both classes are equally distributed (each group consisting of 49 observations). For the variable sex, 26 persons are male and 23 female. Each person was measured twice, resulting in 52 observations for males and 46 for females. We are aware that measuring each person twice introduces over-optimism in the p-values, because the variance in the dataset is smaller with repeated measurements than with independent subjects. However, since the measurements were taken under different conditions (different device and different ‘juice-status’), we

Figure 5. Examples showing the effect of different scaling methods. Top: a metabolite from the dataset with peak identification manual. Bottom: a metabolite from the dataset with peak identification automatic with alignment.
assume that this effect is not large and include them nonetheless, considering this fact in the interpretation. For smoking, there are only four current smokers and five ex-smokers. We merge these two groups to nine (ex-)smokers, which is still a small number for testing and there is a risk that a smoking effect is diluted by the ex-smokers. Here, each person is included twice, analogously to the sex, resulting in 18 measurements for (ex-)smokers and 70 for non-smokers.

Table 2 shows the number of significant metabolites for the three dataset versions (manual, automatic, automatic with alignment). Both, unadjusted $p$-values, and $p$-values adjusted for multiple testing per group variable are reported. For the interpretation, we focus on adjusted $p$-values, and we report unadjusted $p$-values for descriptive purposes. Results are shown for the scale-data (data scaled to mean zero and variance one per device) and scale$^+$-data (data scaled analogously, but shifting the zeros of both devices together). In the manual dataset, that contains more than twice as many variables as the automated dataset, there are always more significant metabolites identified, except for testing for differences between the devices. In the manual dataset, where only very few zeros occur, there is almost no difference between the versions scale and scale$^+$, such that the significant metabolites are exactly the same for both scalings. Overall, the non-parametric Wilcoxon-test often returns more (or equal) numbers of significant metabolites than the Welch- or t-tests for the scale-dataset. For the scale$^+$ dataset it is the other way around.

For juice as grouping variable, 14–44 significant metabolites were identified (after $p$-value-adjustment). For the device as grouping variable, there are no significant variables found, when the scale-dataset is used in combination with the Welch-/$t$-test. This is the expected result, since the mean of the device was set to zero for both devices. For the Wilcoxon-test, where ranks are used instead of the original values, 2–8 significant variables are identified in the scale-dataset. For the scale$^+$-dataset, where the means of both devices are not equal when zeros occur, 18 significant metabolites are identified in the automatic dataset and 7–8 in the automatic with alignment dataset. For sex as grouping variable, no significant variables remain after adjusting the $p$-values, regardless of peak detection and scaling. For smoking, there is one significant metabolite identified in the automatic dataset, two in the automatic with alignment dataset and none in the manual dataset. We interpret these results in the discussion.

5.4. Classification results

For the classification of the observations into the groups with/without juice, we use a 10 fold cross-validation procedure, in order to avoid over-optimistic results. Additionally, each cross-validation is repeated 50 times in order to observe the variability, resulting from different random splits into training and test set. After each replication of a cross-validation, the probabilities for the positive class (here: ‘with juice’) are calculated. In order to obtain summarized performance measures over the replications, the predicted probabilities for the positive class from all replications are merged.

Table 3 shows the resulting performance measures for all three datasets without scaling and with the scalings scale and scale$^+$. Regardless of the combination of dataset version or scaling, the classification works almost perfectly with AUC values not lower than 0.999

Figure 6. PCA for aligned breath data with mean 0 and variance 1 scaling (left, scale) and mean 0 and variance 1 scaling with zeros treated differently (right, scale$^+$).
Table 2. Number of significant metabolites in automatic dataset (top), automatic with alignment dataset (center) and manual dataset (bottom). For the grouping variables marked with ‘*’, a paired t-test was performed instead of a Welch-test.

| Grouping | Scale | Welch-/t-test | Wilcoxon | Scale | Welch-/t-test | Wilcoxon |
|----------|-------|---------------|----------|-------|---------------|----------|
|          | Adj.  | Unadj.        | Adj.     | Unadj. | Adj.          | Unadj.   |
| Juice*   | 16    | (20)          | 18       | (24)   | 16            | (21)     | 14        | (22)     |
| Device*  | 0     | (0)           | 7        | (13)   | 18            | (25)     | 18        | (23)     |
| Sex      | 0     | (1)           | 0        | (2)    | 0             | (0)      | 0         | (2)      |
| Smoke    | 1     | (4)           | 0        | (2)    | 1             | (6)      | 0         | (2)      |

| Grouping | Scale | Welch-/t-test | Wilcoxon | Scale | Welch-/t-test | Wilcoxon |
|----------|-------|---------------|----------|-------|---------------|----------|
|          | Adj.  | Unadj.        | Adj.     | Unadj. | Adj.          | Unadj.   |
| Juice*   | 15    | (19)          | 15       | (21)   | 15            | (20)     | 14        | (20)     |
| Device*  | 0     | (0)           | 8        | (14)   | 8             | (14)     | 7         | (14)     |
| Sex      | 0     | (2)           | 0        | (1)    | 0             | (1)      | 0         | (1)      |
| Smoke    | 2     | (5)           | 0        | (3)    | 2             | (5)      | 0         | (3)      |

| Grouping | Scale | Welch-/t-test | Wilcoxon | Scale | Welch-/t-test | Wilcoxon |
|----------|-------|---------------|----------|-------|---------------|----------|
|          | Adj.  | Unadj.        | Adj.     | Unadj. | Adj.          | Unadj.   |
| Juice*   | 39    | (59)          | 44       | (60)   | 39            | (59)     | 44        | (60)     |
| Device*  | 0     | (0)           | 2        | (6)    | 0             | (0)      | 2         | (6)      |
| Sex      | 0     | (6)           | 0        | (5)    | 0             | (6)      | 0         | (5)      |
| Smoke    | 0     | (19)          | 0        | (6)    | 0             | (19)     | 0         | (6)      |

(AUC = 1 would be the perfect result). The mean misclassification error (MMCE) equals classifying one of the 98 observations wrong (MMCE = 0.01) for the dataset automatic. This value is a bit higher for the dataset manual and lower for the dataset automatic with alignment. The other performance measures also reach almost perfect values. Overall, the methods work equally well on this classification task.

Figure 7 shows the three most important variables in classification for the datasets automatic, automatic with alignment, and manual, when scale or scale\* is used (the three variables with the largest importance variables are the same for both scalings). For each variable, the boxplots for both groups with/without juice are displayed. In all cases, the observations with juice realize larger values than without juice, on average. There are even two metabolites (P23 and P25) for the automatic data with alignment and two for the manual dataset (P91 and P41), that perfectly separate both groups.

In figure 8 we show that the important peaks in classification refer to similar positions in the raw measurements across the peak detection methods. The positions of the five most important peaks are shown as crosses. Depending on the used scaling, the top five peaks are not always the same for one peak detection method, but there is only little variation in the order. Therefore, the five most frequent among the top five of each scaling were considered the five most important ones for a peak detection method. As can be seen, the areas of the five most important peaks in the manual peak detection also contain peaks of the automatic peak detection with alignment. Two pairs of peaks are merged by the alignment step such that two additional peaks are in the top five for the automatic peak detection with alignment. Note that the averaged raw measurement of the automatic detection with alignment looks different, because the alignment was also applied to the raw measurements. In the corresponding image (center), there is just one peak visible where it looks like two in the two other images. Thus, it is plausible that the manual peaks are redundant here and the five positions refer to only three true peaks. These three peaks are among the most important ones for all three peak detection methods. The automatic peak detection without alignment and the automatic peak detection with alignment differ in the positions of two additional peaks.

As a demonstration, we show the effect on the classification, when the differences between the devices are entirely neglected. Assume performing a study in a laboratory with one device (device A). We train a RF on the corresponding dataset with automatic peak identification (without alignment, since we only have
one device). Afterwards, we obtain data from another laboratory that tested approximately the same number of people on their only device (device B). We use this data as a test set and apply our training model from device A on the observations from device B in order to predict the classes of all observations from device B.

Applying this approach to our dataset, the misclassification rate MMCE for device B is 0.51, i.e. about 50% of the predictions are wrong. Figure 9 shows the estimated probabilities for the observations, divided by training set (device A) and test set (device B) as well as their true class (with/without juice). The probabilities in the test set are much lower, such that all observations are classified as ‘without juice’.

The reason of the problem is illustrated in figure 10. Three of the four most important variables, according to the variable importance of the RF, were only identified on device A. Since they separate the
groups perfectly, they were used in the RF model that was only trained data from device A. When applying this model to the data from device B, all observations are zero for these metabolites, consequently indicating that the correct class is 'without juice'. Because 24 of the 49 observations measured on device B actually are in this class, the misclassification rate is \((1 - 24/49) \approx 0.51\), the result reported above.

In Table 4 we show the MMCEs for all combinations of the automatic, automatic with alignment and manual datasets, without scaling, with \(\text{scale}\) and with \(\text{scale}^+\), and also switching the 'training device' (such that the training set is from device B and the test set from device A). It can be seen, that the drastic scenario described above only holds for the automatic peak detection, when no alignment is performed. Using the datasets automatic with alignment or manual peak detection.

### Table 4. MMCE when training was performed on the data from one device and testing on the data from the other device.

| Training device | Peak detection | None | Scale | Scale$^+$ |
|-----------------|----------------|------|-------|---------|
| A               | without alignment | 0.510 | 0.490 | 0.510   |
|                 | with alignment   | 0.041 | 0.041 | 0.041   |
|                 | manual           | 0.020 | 0.041 | 0.041   |
| B               | without alignment | 0.490 | 0.510 | 0.490   |
|                 | with alignment   | 0.000 | 0.000 | 0.000   |
|                 | manual           | 0.020 | 0.020 | 0.020   |

Figure 8. Detail from the average of all 345 raw measurements. The positions of the five most important peaks for the three datasets are displayed as crosses, the remaining peaks as circles. Top: automatic (without alignment). Middle: automatic with alignment. Bottom: manual.
detection, then there is no problem. This shows the drastic improvement of the automatic peak detection when the alignment step is inserted.

6. Discussion

6.1. Alignment of peak positions
For broad applicability of classification models from MCC-IMS-measurements, differences between devices need to be considered. Here, we applied a simple linear regression to shift the peaks from one device to the positions from another device. As reference, we used database information about the peak positions from synthetic gas compositions on both devices. For further studies we recommend to even measure the exact same gas composition under the same conditions as for all other measurements in the study. In this way biases regarding the time of the measurement, the room air,
etc, can be avoided. Also, when applying our linear regression, a reference device must be specified. Here, we arbitrarily chose device A as reference. When many devices are used, more attention needs to be paid to the choice of the reference positions, also regarding the compatibility with databases that help identifying the chemical compounds that are represented by peak positions.

6.2. Interpretation of scalings
For scaling in order to diminish the effect of the device after the automatic peak detection, we focused on \textit{scale} (scaling all data to mean zero and variance one) and \textit{scale}+ (treating all values that are exactly zero differently, shifting the zeros for both devices by the same value). Since the classification problem (with/without juice) was easy, no big differences for classification accuracy were observed. However, the differences can descriptively be evaluated by the examples of the metabolite distributions in figure 5 and the PCA plots in figures 4 and 6. For the automatic peak detection data, \textit{scale}+ looks more suitable for scaling, see figure 5 (bottom), but the effect of the device cannot entirely be compensated. Therefore, the device still accounts for a large proportion of the variance in the data, see figure 6.

On the other hand, \textit{scale} is not as intuitive to interpret, when zeros occur, but it scales the data strictly to mean zero and variance one. In our experiment, both devices measured about the same numbers of subjects with and without juice (device A: 25 with juice, device B: 24 with juice). Therefore, we can assume that the observations for both devices are comparable and both devices measure the same effects on average. Thus, the number of zeros for both devices should be approximately the same, if both devices are comparable. In reality, however, there are metabolites that are mainly measured only on one device, for different reasons. First, it is possible, that the positions of the metabolites for both devices are so far apart from each other that they are not detected as the same component and second, it is possible, that the metabolite is actually only present in one device, for example due to different materials built into the devices. In the first case, the device fails to detect a present metabolite, so the intensities should not be interpreted as zero, but as missing values. In the second case, a metabolite is only present in one device, so values on the other device should really be zero. In this case, there is a real difference between the devices, and no scaling is required. Currently, we cannot discriminate between these two scenarios for the metabolites in our study.

In the case that one metabolite takes only zero values for one device, and only non-zeros for the other device, \textit{scale} corresponds to treating zeros as missing values. After the scaling, original zeros are still zero, and original non-zeros jitter around zero. This is comparable to exchanging missing values by the average value of non-missing values. On the other hand, \textit{scale}+ would shift the zero-values such that they are smaller than all scaled non-zero values. This corresponds to interpreting zeros as actually absent metabolites and assuming that zeros could have occurred on the other device as well and that their shifting should have been the same.

In general, the scaling methods can also be applied to adjust for other confounders. However, this requires that the observations in the confounder groups are comparable and do not differ systematically with respect to other variables. In our case, scaling with respect to the device works straightforward because each person was measured once on each device, such that variables like sex and smoking status are equally represented for the devices and for the measurements with/without juice. Therefore, it is legitimate to assume, that the observed metabolite intensities should be comparable on both devices.

The different scalings should be further investigated on different devices to analyze their impact on classification performance. Also, outgassing metabolites should be known for each device, such that they can automatically be eliminated from the dataset before analyzing the breath gas. Further, an additional correction step for the peak clustering could be considered, where implausible results are corrected. When two different relatively close peaks are detected on two devices, where each one is basically only detected on one device, although both devices measure similar subjects, they could be merged. To avoid zeros in the automatic peak detection, an additional step could be added to the peak detection algorithm, which checks at every consensus peak position in each raw measurement, if the intensity values are high enough to assume that there is actually a peak that has just not been detected in the peak picking step.

6.3. Significant metabolites in smoking
After adjusting the \(p\)-values of the univariate tests (Welch-test) for different means (see table 2), there is one metabolite that significantly discriminates between non-smokers and (ex-) smokers in the automatic dataset and two metabolites in the automatic with alignment dataset. The significant metabolites are the same for \textit{scale} and \textit{scale}+ . For the manual dataset, no metabolites are significant after the adjustment.

For the significant metabolites, figure 11 shows their distribution separately for the smoking status, exemplary with \textit{scale}+ scaling. It can be seen that for two of the three metabolites of interest, all observations for the smokers, in case of P14 all except for two, realize the same value (zero before scaling), whereas 20%–30% of the non-smokers realize higher values, causing a significant test result.

However, it must be noted, that these metabolites lack discriminating potential since the majority of non-smokers realize the same value as the (ex-) smokers. Furthermore, it seems counterintuitive that
smokers should stand out by missing metabolites instead of showing additional ones, although this is not biologically impossible since smoking could inhibit metabolic processes of non-smokers.

As a conclusion, the significant metabolites in smoking are most likely to be artifacts, resulting from the use of different devices and from the automated peak detection.

7. Conclusions

We conducted a study including 49 healthy participants with the aim to find and diminish the effect of confounders, simultaneously comparing the results from the currently popular semi-manual peak detection with fully automated peak detection. In order to account for the effect of the device, the automatic peak detection was additionally analyzed in combination with a peak alignment step. As a substitute for different disease groups, we measured study participants twice, once before and once after drinking a glass of orange juice. As examples for confounders we considered the sex and the smoking status of a person as well as the MCC-IMS device that was used. Therefore, we used two devices in the study such that one of the two measurements per person was taken on each device.

Different scalings were applied in order to compensate for the effect of the device. Descriptive analyses showed that scaling to mean zero and variance one works well for the manually detected peaks. Peaks identified with automatic peak identification lead to more intuitive results when an additional scaling step is applied, that treats zero values in the dataset different from the remaining values. The two scalings scale and $scale^+$ were considered. When outliers occur, a robust alternative is suggested.

Univariate tests for location with adjusted p-values identified about 15 significant metabolites for the target variable with/without juice for the automatic and about 40 metabolites for the manual peak detection. The tests for the target variable device are as expected not significant, when $scale$ is used in combination with a $t$-test. For the automatic peak detection 8–18 significant metabolites remain after applying $scale^+$. Testing for differences of location for the sex, no significant metabolites were found, regardless of the used scaling or peak detection. For smoking, 1–2 significant metabolites were identified when applying the
automatic peak detection, but at a closer look they are likely to be artifacts. Note that the number of (ex-)smokers is only nine out of 49 and that a potential effect of smoking on the breath measurements could be diluted by combining smokers and ex-smokers to one group.

The classification task with respect to the groups with/without juice was extremely simple for the classification algorithm RF. Therefore, no big differences could be observed for the different scalings or peak detection methods.

As a demonstration, we showed that in case of using the automatic peak detection without peak position alignment, the classification can fail entirely, when the classification model is built on the data from one device and afterwards applied to the data from the other device. Although this effect was not present when the peak position alignment was performed or when the manual peak detection was used, the risks due to unknown confounders became apparent.

Although in our study we did not find clear differences between men and women as well as between (ex-)smokers and non-smokers, more attention should be paid to possible confounders in the future. Test and control groups should stem from the same population. In case of present confounding factors, appropriate statistical methods should be used.

We showed that the two devices that were used lead to different results. The peak positions of both devices are slightly shifted, causing problems for the automatic algorithm. We further showed that the results clearly improve, when the peak positions are aligned, using a linear regression of a reference gas composition. Different signal intensity levels were compensated for by using scaling methods. Future research is required to solve these problems for more than two devices. For each device the metabolites that are outgassing from the device should be known, together with their corresponding peak positions, since the resulting peaks are potentially specific to the device. Those metabolites should then be eliminated from the dataset.

Acknowledgments

This study was supported by Deutsche Forschungsgemeinschaft (DFG, Germany) within the Collaborative Research Center (Sonderforschungsbereich) SFB 876 ‘Providing Information by Resource-Constrained Analysis’, project TB1. We thank Elias Kuthe, Alexey Egorov, and the corresponding project group for providing software for the automatic peak detection as described in [14].

ORCID iDs

S Horsch © https://orcid.org/0000-0001-7511-6612

References

[1] Cao W and Duan Y 2006 Breath analysis: potential for clinical diagnosis and exposure assessment Clin. Chem. 52 800–11
[2] Fink T, Baumbach J I and Kreuser S 2014 Ion mobility spectrometry in breath research J. Breath Res. 8 027104
[3] Cheng Z J, Warwick G, Yates D H and Thomas P S 2009 An electronic nose in the discrimination of breath from smokers and non-smokers: a model for toxin exposure J. Breath Res. 3 036003
[4] McWilliams A, Beigi P, Srinidhi A, Lam S and MacAulay C E 2015 Sex and smoking status effects on the early detection of early lung cancer in high-risk smokers using an electronic nose IEEE Trans. Biomed. Eng. 62 2044–54
[5] Pourhoseingholi M A, Baghestani A R and Vahedi M 2012 How to control confounding effects by statistical analysis Gastroenterol. Hepatol. Bed Bench 5 79–83
[6] Ernstgard L, Stiggen B, Warholm M and Johanson G 2003 Sex differences in the toxicokinetics of inhaled solvent vapors in humans 2. 2-Propanol Toxicol. Appl. Pharmacol. 193 158–67
[7] Kischkel S, Miekisch W, Sawacki A, Straker E M, Trefz P, Amann A and Schubert J K 2010 Breath biomarkers for lung cancer detection and assessment of smoking related effects—confounding variables, influence of normalization and statistical algorithms Clin. Chem. Acta 411 1637–44
[8] Hauschild A C, Frisch T, Baumbach J I and Baumbach J I 2015 Carotta: revealing hidden confounder markers in metabolic breath profiles Metabolites 5 344–63
[9] Baumbach J I 2009 Ion mobility spectrometry coupled with multi-capillary columns for metabolic profiling of human breath J. Breath Res. 3 034001
[10] Horsch S, Kopczynski D, Kuthe E, Baumbach J I, Rahmann S and Rahnenführer J 2017 A detailed comparison of analysis processes for MCC-IMS data in disease classification—automated methods can replace manual peak annotations PLoS One 12 e0184521
[11] Cumaras R, Schneider T, Favrod P, Figueras E, Gracia I, Maddula S and Baumbach J I 2012 Stability and alignment of MCC/IMS devices Int. J. Ion Mobil. Spec. 15 41–6
[12] Hauschild A C, Baumbach J I and Baumbach J I 2012 Integrated statistical learning of metabolic ion mobility spectrometry profiles for pulmonary disease identification Genet. Mol. Res. 11 2733–44
[13] R Core Team 2018 R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing Vienna, Austria
[14] Egorov A et al 2014 Ressourcenbeschränkte Analyse von Ionenmobilitätspektren mit dem Raspberry Pi Technical Report Faculty of Computer Science, TU Dortmund
[15] Hastie T, Tibshirani R and Friedman J 2009 The Elements of Statistical Learning (Springer Series in Statistics) 2nd edn (New York: Springer)