Feasibility study on exhaled-breath analysis by untargeted Selected-Ion Flow-Tube Mass Spectrometry in children with cystic fibrosis, asthma, and healthy controls: comparison of data pretreatment and classification techniques.

Segers, Karen; Slosse, Amorn; Viaene, Johan; Bannier, Michiel A.G.E.; Van de Kant, Kim D.G.; Dompeling, Edward; Van Eeckhaut, Ann; Vercammen, Joeri; Vander Heyden, Yvan

Published in:
Talanta

DOI:
10.1016/j.talanta.2021.122080

Publication date:
2021

Feasibility study on exhaled-breath analysis by untargeted Selected-Ion Flow-Tube Mass Spectrometry in children with cystic fibrosis, asthma, and healthy controls: comparison of data pretreatment and classification techniques

Karen Segers[a,b], Amorn Slosse[a], Johan Viaene[a], Michiel A. G. E. Bannier[c], Kim D. G. Van de Kant[c], Edward Dompeling[c], Ann Van Eeckhaut[b], Joeri Vercammen[d,e], Yvan Vander Heyden[a]*

Double family name: Vander Heyden

Van Eeckhaut

Triple family name: Van de Kant

[a] Department of Analytical Chemistry, Applied Chemometrics and Molecular Modelling, Vrije Universiteit Brussel (VUB), Laarbeeklaan 103, 1090 Brussels, Belgium

[b] Department of Pharmaceutical Chemistry, Drug Analysis and Drug Information, Center for Neurosciences (C4N), Vrije Universiteit Brussel (VUB), Laarbeeklaan 103, 1090 Brussels, Belgium

[c] Department of Paediatric Respiratory Medicine, School for Public Health and Primary Care, Maastricht University Medical Centre+, Maastricht, The Netherlands

[d] Interscience Expert Center (IS-X), Avenue Jean-Etienne Lenoir 2, 1348 Louvain-la-Neuve, Belgium

[e] Industrial Catalysis and Adsorption Technology (INCAT), Faculty of Engineering and Architecture, Ghent University, Valentin Vaerwyckweg 1, 9000 Ghent, Belgium

e-mail addresses authors :

Karen Segers : Karen.Segers@vub.be

Amorn Slosse : Amorn.Slosse@vub.be

Johan Viaene : Johan.Viaene@vub.be

Yvan Vander Heyden : yvanvdh@vub.be

Joeri Vercammen : J.Vercammen@is-x.com

Michiel Bannier : michiel.bannier@mumc.nl

Kim Van de Kant : kim.vande.kant@mumc.nl

Edward Dompeling : edward.dompeling@mumc.nl

Ann Van Eeckhaut: aveeckha@vub.be

*** Corresponding author:**

Vander Heyden Yvan

Department of Analytical Chemistry, Applied Chemometrics and Molecular Modelling, Center for Pharmaceutical Research (CePhaR), Vrije Universiteit Brussel (VUB)

Laarbeeklaan 103

1090 Brussels

Belgium

yvanvdh@vub.be

Abstract

Selected-Ion Flow-Tube Mass Spectrometry (SIFT-MS) has been applied in a clinical context as diagnostic tool for breath samples using target biomarkers. Exhaled breath sampling is non-invasive and therefore much more patient friendly compared to bronchoscopy, which is the golden standard for evaluating airway inflammation. In the actual pilot study, 55 exhaled breath samples of children with asthma, cystic-fibrosis and healthy individuals were included. Rather than focusing on the analysis of target biomarkers or on the identification of biomarkers, different data analysis strategies, including a variety of pretreatment, classification and discrimination techniques, are evaluated regarding their capacity to distinguish the three classes based on subtle differences in their full scan SIFT-MS spectra. Proper data-analysis strategies are required because these full scan spectra contain much external, i.e. unwanted, variation. Each SIFT-MS analysis generates three spectra resulting from ion-molecule reactions of analyte molecules with H_3O^+, NO^+ and O_2^+. Models were built with Linear Discriminant Analysis, Quadratic Discriminant Analysis, Soft Independent Modelling of Class Analogy, Partial Least Squares - Discriminant Analysis, K-nearest Neighbours, and Classification and Regression Trees. Perfect models, concerning overall sensitivity and specificity (100% for both) were found using Direct Orthogonal Signal Correction (DOSC) pretreatment. Given the uncertainty related to the classification models associated with DOSC pretreatments (i.e. good classification found also for random classes), other models are built applying other preprocessing approaches. A Partial Least Squares - Discriminant Analysis model with a combined pre-processing method considering single value imputation results in 100% sensitivity and specificity for calibration, but was less good predictive. Pareto scaling prior to Quadratic Discriminant Analysis resulted in 41/55 correctly classified samples for calibration and 34/55 for cross-validation. In future, the uncertainty with DOSC and the applicability of the promising preprocessing methods and models must be further studied applying a larger representative data set with a more extensive number of samples for each class. Nevertheless, this pilot study showed already some potential for the untargeted SIFT-MS application as a rapid pattern-recognition technique, useful in the diagnosis of clinical breath samples.

Keywords: Exhaled breath analysis, Selected-Ion Flow-Tube Mass Spectrometry, Principal component analysis, Classification and discrimination, Data Preprocessing techniques

Abbreviations

ACT	Asthma control test
CART	Classification and regression trees
d_c	Dissimilarities
DFs	Discriminant functions
DOSC	Direct orthogonal signal correction
FeNO	Exhaled nitric oxide
FEV_1	Forced expiratory volume in 1 second
FN	False negatives
FP	False positives
ICS	Inhaled corticosteroids
KNN	K-nearest neighbours
LABA	Long-acting β2-agonists
m/z	mass-to-charge ratio
PC	Principal component
PCA	Principal component analysis
PCA-LDA	Principal component analysis - linear discriminant analysis
PCA-QDA	Principal component analysis - quadratic discriminant analysis
PLS-DA	Partial least squares - discriminant analysis
PQN	Probabilistic quotient normalisation
r	Pearson correlation coefficient
SD	Standard deviation
SIFT-MS	Selected-Ion Flow-Tube Mass Spectrometry
SIM	Selected ion monitoring
SIMCA	Soft independent modelling by class analogy
SNV	Standard normal variate
TN	True negatives

TP	True positives
VIP	Variable Importance in Projection
VC	Volatile compound
VOC	Volatile organic compound

1. Introduction

The evaluation of airway inflammation in lung diseases is typically carried out by means of bronchoscopy [1, 2]. Since it is an invasive technique, alternative analysis techniques have been proposed as presented in the review article by Bannier et al. [2]. In that paper, the analysis of exhaled breath volatiles for evaluating lung diseases is discussed as well. Besides lung diseases, exhaled breath analysis has also been proposed to monitor other diseases, such as various cancers, metabolic disorders, hepatitis and gastroenteric diseases [1, 3-7]. In all cases, diagnosis is based on the presence of specific volatile biomarkers in the exhaled breath [1, 3, 6, 8].

Several instrumental techniques have been proposed to analyse volatiles in exhaled breath. Most acknowledged approaches focus on the analysis of volatile organic components (VOCs) and use thermal desorption analysis in combination with high resolution techniques, such as GC-MS [7, 9]. Generally, samples are collected by means of direct exhalation into a suitable polymeric sampling bag. Polymeric bags are relatively cheap and very convenient but susceptible to diffusion of permanent gases through the wall and/or elevated blank levels. More information about bag materials for breath samples can be found in [10]. Very soon after sampling, the exhaled breath is transferred to a thermal desorption tube that is packed with an appropriate adsorbent (or combination of adsorbents). Thermal desorption tubes are very easy to handle and permit sample storage over prolonged periods of time, making it more in line with typical GC-MS turnaround delays [9].

GC-MS in full scan mode is particularly well-suited for biomarker discovery because of its capacity to deconvolute and identify individual chromatographic peaks based on their (high resolution) mass spectra [11, 12]. Nonetheless, GC-MS is far too complicated to be employed as a dedicated point-of-care device by non-specialists in a clinical context, such as for direct exhaled breath analysis. This opportunity gap is elegantly bridged by chemical sensor arrays that hold the promise of fast, sensitive and selective detection of the biomarkers, earlier identified by means of GC-MS. Although these arrays show promising results, the applied methodology suffers from some severe shortcomings [6]. Most importantly, it does not account for analytical bias towards small polar analytes, reactive components and inorganic volatiles that result from the use of thermal desorption GC-MS or that might be present in the humid exhaled breath [9].

In that respect Selected-Ion Flow-Tube Mass Spectrometry (SIFT-MS) is better, giving rise to a more comprehensive analysis of exhaled breath. SIFT-MS is a type of direct mass spectrometry that allows sensitive and selective detection of volatile organic and inorganic compounds in gaseous samples, without the need for complicated sample preparation procedures that might affect compound recovery [4, 5, 13]. The basic operational

principles of the technique are presented in Figure 1. Briefly, it uses soft chemical reactions that occur between multiple precursor ions (H_3O^+, NO^+ and O_2^+) that are generated *in situ* in the ionization region of the instrument and are introduced one-by-one into the reaction chamber or flow tube using a short upstream quadrupole. As they enter the flow tube, precursor ions are thermalized by means of a high flow of helium carrier gas. Afterwards, they react rapidly with the sample molecules which are introduced in the flow tube. Since each precursor ion is able to react differently with isobaric components, a degree of selectivity is obtained that outperforms sensor arrays, particularly when complex samples, such as exhaled breath, are involved [4, 13].

In general, SIFT-MS is used in targeted or selected ion monitoring (SIM) mode, which means that the components of interest are known beforehand. For instance, for the fast quantification of components that were earlier identified from, for example, GC-MS [14, 15]. Alternatively, SIFT-MS in full scan mode is a more delicate approach since the entire chemical identity of a particular sample is recorded in a minute span of time. It is applied less frequently, because of the presence of not-disease-related ("irrelevant") variation in the breath-sample composition, making the interpretation of the results more complicated. The variation is related to exogenous exposures, such as food intake or medication. Those exposures may interact with the volatile compound (VC) composition [16, 17]. The goal of this feasibility study is interpreting the abstract nature of the full scan data, which requires specific data analysis and visualization procedures that are able to extract the relevant information contained within the full scan spectra. Those data analysis strategies follow often a trial-and-error principle, resulting in an enormous potential workload for the scientist. Therefore, this pilot study aims demonstrating which data-analysis strategies are valuable for further consideration in untargeted SIFT-MS profiling of breath samples for rapid pattern-based screening. Certain data-analysis approaches applied on the SIFT-MS data have already shown their usefulness for various applications, such as classifying olive and Argan oils by means of headspace aroma analysis [18], and in a clinical context [4].

In the present feasibility study, the full scan SIFT-MS spectra of exhaled breath samples from 55 children, i.e. 20 healthy, 22 asthmatic and 13 with cystic fibrosis, were analysed. The goal of our study was not to identify target biomarkers, but to investigate which data analysis strategy allows a maximal distinction between the groups of children. Additionally, principal masses were associated with biomarkers previously reported in the literature.

Unsupervised data analysis was performed using Principal Component Analysis (PCA) with visual evaluation of score and loading plots. Supervised analysis consist of K-nearest Neighbours (KNN), Classification and Regression Trees (CART), PCA - Linear Discriminant Analysis (PCA-LDA), PCA - Quadratic Discriminant Analysis (PCA-QDA), Soft Independent Modelling by Class Analogy (SIMCA) and Partial Least Squares - Discriminant Analysis

(PLS-DA). The quality of the models was evaluated by the calibration and cross-validation errors, the % overall sensitivity, % overall specificity and % model efficiency [19, 20].

2. Theory

2.1. Data preprocessing

Spectra often contain noise or variables that are irrelevant for the studied classification problem. To reduce the undesired data variation, preprocessing techniques are applied. They remove for instance noise. More specific, treatments such as variable reduction or elimination remove irrelevant variables, while relevant information (for classification) will be maintained [20-22]. Variable selection methods, on the other hand, selects the relevant variables. Additionally, it may be important for all variables to be comparable in magnitude and to have similar ranges [23].

The data matrix **X** is an n x p matrix, with n the number of samples and p the number of variables. These variables are in this case study the *m/z* ratios of the formed product ions, resulting from the reaction between a precursor ion (H_3O^+, NO^+ and O_2^+) and compounds occurring in exhaled breath. The measured response for each variable is the signal intensity.

Nineteen preprocessing approaches were performed on **X**; 1) column centering, 2) Pareto scaling, 3) Dong's Algorithm to remove non-significant variables, 4) Centering after Dong's Algorithm, 5) Autoscaling after Dong's Algorithm, 6) Pareto scaling after Dong's Algorithm, 7) Normalisation by the norm and column centering after Dong's Algorithm, 8) Probabilistic quotient normalisation (PQN) after Dong's Algorithm, 9) Standard Normal Variate (SNV) and centering after Dong's Algorithm,10) Direct Orthogonal Signal Correction (DOSC) after Dong's Algorithm, 11) DOSC on the raw data, 12) Single value imputation to replace the zero values by the mean followed by normalisation by the norm, 13) Single value imputation to replace the zero values by the mean followed by PQN normalisation, 14) Single value imputation to replace the zero values by the median followed by normalisation by the norm, 15) Single value imputation to replace the zero values by the median followed by PQN normalisation. The preprocessing results from 12) to 15) were log transformed and autoscaled. These last data preprocessing approaches (12-15) were also applied in combination with Dong's Algorithm (16-19). Approaches 12-15 were found to be suitable in untargeted full-scan SIFT-MS analyses, as a diagnostic tool, for asthma phenotyping [24].

Column centering subtracts from each column element the respective column average [25, 26]. Autoscaling, also called column standardization, is column centering followed by division by the column standard deviation [27]. This normalisation gives each variable an equal weight (same average, same standard deviation) [25, 26]. Another often applied

scaling method is Pareto scaling, which is similar to autoscaling but instead of the standard deviation its square root is used [28]. Normalisation by dividing each row element by its norm (i.e. the square root of the sum of all squared elements in that row) [23] was also performed as well as PQN, where each variable is normalized by the median quotient. SNV is a normalisation with row centering and row scaling [22, 26].

DOSC removes the information that is not orthogonal to the class information [29]. The **X** matrix is corrected for variations that are not orthogonal with **y** (classes of the samples).

To remove the non-significant (noise) variables, Dong's algorithm was applied as a variable reduction technique [30, 31].

Classification and discrimination techniques were performed on the 19 preprocessed **X** matrices and on the raw data matrix. First, unsupervised classification was visually evaluated on PCA score plots.

2.2. Unsupervised exploratory analysis

PCA reduces the number of original variables by creating new (latent) variables, principal components (PCs), which are linear combinations of the original variables. PCA allows visualizing the information and variation included in **X**. The variation is presented in the PCs, with the first PC (PC1) representing the largest variation. The second PC (PC2) is orthogonal to PC1, describes most of the remaining variation (less than PC 1) and is defined in the direction of the largest remaining variance not explained by PC1. The coordinates of the projection of the samples on the new variables (PCs) are called scores, which can be represented in a score plot. The scores are weighted linear combinations of the original variables. A score plot may be a one-, two- or three-dimensional plot representing the score(s) of the samples on one, two or three PCs. It reflects information about similarities and differences between the samples. The weights of the original variables in the scores are called loadings. A loading plot shows information about the original variables, for instance, their correlation [22, 26, 32].

2.3. Supervised classification and discrimination analysis

In supervised classification and discrimination techniques, the information present in matrix **X** is, most often, related to an n x 1 response vector **y**, representing the classes of the samples [33]. In this study, different techniques, such as KNN, CART, PCA-LDA, PCA-QDA, SIMCA, and PLS-DA, are used to model **y** as a function of **X**.

Classification techniques describe one class at the time. These techniques model an enclosed class space. The shape of this space is characteristic for the classification technique applied. If two or more classes are modelled, the obtained spaces may overlap, resulting in the possibility that a sample is compatible with more than one class.

Additionally, a part of the global multidimensional domain will not be included in the class spaces. This may result in samples that do not belong to any of the modelled classes. Discriminant methods require at least two classes. A delimiter is described that divides the global domain in a number of regions, each assigned to one class. The type of delimiter is specific for a given discriminant method. In the latter methods, class areas will never overlap and there is no possibility of non-assignment of samples [20].

The predictive ability of the obtained model was evaluated by venetian blind cross-validation or by using an independent test set [20]. In venetian blind cross-validation the samples in the cross-validation groups are selected regularly spread across the matrix [19]. A 5-groups venetian blinds cross-validation is applied ensuring that all classes are present in each test set. This approach of validation results in a lower risk of overestimating the predictive power of a given model, which is more often the case with leave-one-out cross-validation.

The quality evaluation of the models was based on their overall specificity, sensitivity, model efficiency and number of not-assigned samples. First, each class *i* was individually considered and the samples were predicted as true positives (TP), true negatives (TN), false negatives (FN) or false positives (FP) as shown in Table 1. TP are the class members assigned to the considered class, while TN are the non-class members not assigned to that class. Furthermore, FN are the considered class members that were not assigned to that class and FP are the non-class members assigned to the considered class. An illustration in perspective of class A is given in Table 1. Samples belonging to class A and predicted as class A are TP. Samples belonging to class A and predicted as a class B/C member are FN. Furthermore, samples belonging to class B/C and predicted as class B/C are TN, while FP are the samples belonging to class B/C and predicted as class A members. The specificity, sensitivity, model efficiency, precision and number of not-assigned samples were first calculated for each class *i* (*i*=3) for each model, according to Eqs. (3) - (6).

$$\%\ \text{specificity}_i = \frac{TN_i}{(TN_i + FP_i)} \cdot 100 \qquad (3)$$

$$\%\ \text{sensitivity}_i = \frac{TP_i}{(TP_i + FN_i)} \cdot 100 \qquad (4)$$

$$\%\ \text{model efficiency} = \sqrt{\%\ \text{sensitivity}_i \cdot \%\ \text{specificity}_i} \qquad (5)$$

$$\%\ \text{precision}_i = \frac{TN_i}{(TP_i + FP_i)} \cdot 100 \qquad (6)$$

Sensitivity reflects the ability of the model to correctly recognize samples belonging to a class, where specificity is the ability of the model to reject samples that are not belonging to that class.

The model efficiency is expected to be high, i.e. none or few samples are incorrectly classified. To get an idea about the total correct classification ability of the model, the precision is evaluated. If the precision is 100 %, it means that all samples are correctly assigned to their class.

Subsequently, to evaluate the models globally, the overall specificity, sensitivity, model efficiency and number of not-assigned samples were determined, based on the individual class parameters, according to Eqs. (7) - (9).

$$\% \text{ overall specificity} = \frac{\sum_{i=1}^{3} \% \text{ specificity}_i \,.\, n_i}{\sum_{i=1}^{3} n_i} \quad (7)$$

$$\% \text{ model efficiency} = \frac{\sum_{i=1}^{3} \% \text{ model efficiency}_i}{3} \quad (8)$$

$$\% \text{ overall sensitivity} = \frac{\sum_{i=1}^{3} TP_i}{\sum_{i=1}^{3} (TP_i + FN_i)} . 100 \quad (9)$$

where n_i is the number of samples in class *i* and number 3 stands for the number of classes.

The % overall sensitivity refers to the ability of the model to predict the correct class.

2.3.1. K- nearest Neighbours (KNN)

KNN is a non-linear supervised technique for classification and regression [20]. It is based on the distance or proximity between samples [3, 34]. The input value for a new sample in a KNN approach is its distance to a measured calibration sample neighbour or the average distance when classification is based on more than one neighbour [35]. When low correlation between the **X** variables occurs, the Euclidean distance is frequently used as measure [34].

Another parameter often applied to express the similarity between neighbours is the Pearson correlation coefficient (r). Dissimilarities (d_c) are defined as d_C= 1-|r| and are similar to the Euclidean distance, i.e. they are low for similar samples [35].

In KNN, first, the best number of neighbours has to be defined, for instance, based on the error of cross validation. The most simple method is when only one neighbour is considered for classification, which is often used when the number of training samples is large and in the absence of outliers. In the presence of outliers some nearest neighbours have to be taken into account. The appropriate number of neighbours K is usually less than 10 [36]. The number of neighbours included will influence the method performance [37, 38].

After determining the proper number of neighbours, new samples can be classified. Prediction of a new sample is based on the category membership of most of its K nearest neighbours [3, 39].

2.3.2. Classification and Regression Trees (CART)

CART is a nonparametric technique, applied for exploratory analysis, regression and classification. This regression/classification technique can be used with both categorical and continuous responses [40].

The building of a tree is based on a binary recursive partitioning of the data. The term ''binary'' implies that each group of samples, represented by a ''node'' in a decision tree, is split into two groups [41]. The separation into child nodes is based on splitting criteria [42].

The classification tree is built by sub-dividing the root or parent node, containing all samples, in two child nodes based on a split value for one of the variables present in the **X** matrix. Each parent node results in two child nodes and each child node may split further in sub-nodes. Nodes that are not split anymore are called terminal nodes [22, 27, 42].

The building of a CART-model consists of three steps. First, an over-large tree is built using recursive partitioning. In this first tree only pure or homogeneous terminal nodes are present. In the second step, branches of the over-large tree are cut to obtain smaller trees, improving the predictive ability without losing accuracy. This second step is called pruning. The last step is to select the optimal tree based on its predictive ability [41, 42].

2.3.3. Linear and quadratic discriminant analysis

PCA-LDA and PCA-QDA are both parametric methods, which assume a Gaussian distribution of the data in the classes [39, 41]. The methods work properly when all classes are strictly homogeneous [22].

The technique reduces the number of variables by constructing latent variables from the numerous original ones, and searches for a maximal discrimination between the classes. This is done by making a linear combination of the original variables that maximizes the between-class variance relative to the within-class variances [3, 22, 43]. The linear combinations are called discriminant functions (DFs). In PCA-QDA the DFs are quadratic [22, 27]. The maximal number of DFs is equal to the number of classes minus 1 [22]. Here, for a 3-class discrimination, maximally 2 DFs are defined.

The limitation of LDA is that the number of variables has to be lower than the number of samples. QDA requires that the number of variables is lower than the number of objects in the smallest class. For the SIFT-MS spectra these requirements are not fulfilled because of the relatively high number of variables registered. These dimensionality problems can be solved by reducing the number of variables with, for instance, PCA prior to LDA and QDA, or with stepwise regression [3, 21, 22, 27, 33, 41]. In this study, PCA is used.

The optimal model complexity for PCA-QDA is often determined by cross-validation. The % correct classification rate was calculated for models with different complexities. The complexity is optimal when the model results in the highest correct classification rate for predicted samples [41]. For PCA-LDA and PCA-QDA, the model complexity was optimised by venetian blinds cross-validation with 5 groups.

2.3.4. Soft independent modelling by class analogy (SIMCA)

SIMCA is a distance-based technique, as KNN [20]. First, PCA models are created for each class individually. The optimal number of PCs is determined independently for each class based on % model efficiency, sensitivity and specificity, in order to have a predefined percentage of explained cumulative variance per class [3, 22].

Consequently, for each class, a model is built in a hyperspace with a number of dimensions equal to the selected number of PCs [3]. For each class, a closed space is constructed at a given level of significance. Since the PCs are orthogonal the space will have the shape of a segment (one PC applied), a rectangle (two PCs) or parallelepiped or hyper-parallelepiped (three or more PCs) form [20]. Samples may be assigned to a specific class based on their shortest distance to that class. This approach results in samples that are assigned to only one class [20, 22].

Samples may also be assigned to a given class based on their global distance to the centre of the respective class. When the global distance does not exceed a given threshold, the global SIMCA distance, the sample is considered to belong to this class. The global threshold is found by increasing the threshold for each class maximizing sensitivity and specificity. This approach is used in our study. Samples were not assigned to a class when the distances exceeded the threshold. Samples may be assigned to more than one class, when the sample distance is below the thresholds of different classes [22]. This often occurs when class spaces have some overlap.

Besides defining boundaries for each class, SIMCA may also be used as an alternative discriminant technique. Then, a delimiter is calculated corresponding to the locus of points with the same distance from the models of at least two classes.

2.3.5. Partial-least-squares discriminant analysis (PLS-DA)

PLS-DA is a linear and parametric classification method. The linear model uses latent variables [25], which describe a maximal covariance between the (spectral) variables and the response [19, 41]. The selection of the best number of latent variables is based on cross-validation results.

The responses in PLS-DA are qualitative, discrete and coded in a vector with numbers 0 and 1, where 1 refers to belonging to a class and 0 not. When three classes are present,

each class is modelled once relative to the rest, applying three vectors with labels (1,0,0), (0,1,0) and (0,0,1), respectively [41]. PLS-DA can be used in different approaches. In a first, samples are classified in one of the three classes based on probability. The predicted value is around 0 or 1. When the value is closer to 0, the sample does not belong to the considered class. Samples are always classified to a class [19].

The second approach, which is applied in this study, uses a threshold for each class. Consequently, a given sample, with a value above the threshold, is considered to belong to the specific class [19, 41]. When the value is lower, the sample is not assigned to that class. The procedure is repeated for every model. Samples not assigned to any class or indicated to several classes are defined as 'not classified'. This may occasionally result in a classification with a high number of not-assigned samples [19].

3. Experimental

3.1. Sample collection

In total, for this pilot study, 55 samples were collected from children at the Maastricht University Medical Centre+ (MUMC+) hospital (Maastricht, The Netherlands) over a period of 6 months. These included 20 children with asthma (average age ± SD: 12.7 ± 3.1 years), 13 children with cystic fibrosis (14.4 ± 4.2), and 22 healthy controls (9.7 ± 2.0). More subject characteristics are shown in Table 2. Written informed consent was obtained from all subjects. The study was approved by the Medical Ethical Committee of the Maastricht University Medical Centre+. All samples were collected in 1 L Tedlar bags with polypropylene valve and septum fitting (Interscience, Breda, The Netherlands). All children were instructed to refrain from: 1) eating and drinking at least 2 hours before testing, with the exception of water, 2) chewing gum or brushing teeth at least 2 hours before testing, 3) exercise at least 1 hour before testing, 4) use of inhalation medication at least 3 hours before testing. Exclusion criteria for this pilot study were a recent course of prednisone or antibiotics within one month before testing (maintenance antibiotics for CF excepted), (second-hand) smoking, and an extra-pulmonary chronic inflammatory disease (e.g. inflammatory bowel disease, rheumatic disease). Finally, all measurements were executed in one room and at the same environmental conditions, e.g. changes in room temperature and humidity were kept to a minimum.
The filled Tedlar bags were transported to Interscience (Breda, The Netherlands) where the breath samples were immediately analysed by SIFT-MS upon arrival.

3.2. SIFT-MS

The Tedlar bag contents were introduced into a Voice200® ultra SIFT-MS instrument (Syft Technologies, Christchurch, New Zealand) at a constant flow rate of 20 mL min^{-1} using the instrument's high vacuum in combination with a fixed restriction installed at the instrument

inlet. Full scan MS spectra of H_3O^+, NO^+ and O_2^+ were recorded between 15 and 250 *m/z* at unit resolution for each precursor; the dwell time was 100 ms per mass at three data points. Instrument calibration was performed on a daily basis by measuring a certified gas cylinder containing the following compounds: benzene ($C_6H_6^+$ [O_2^+], m/z = 78), ethylene ($C_2H_4^+$ [O_2^+], m/z = 28), hexafluorobenzene ($C_6F_6^+$ [O_2^+], m/z = 186), isobutene ($C_4H_8^+$ [NO^+], m/z = 57), octofluorotoluene ($C_7F_8^+$ [O_2^+], m/z = 236), tetrafluorobenzene ($C_6F_4H_2^+$ [O_2^+], m/z = 150) and toluene ($C_7H_8.H^+$ [H_3O^+], m/z = 93; $C_7H_8^+$ [NO^+], m/z = 92 and $C_7H_8^+$ [O_2^+], m/z = 92). For each sample, relative humidity was estimated by summing the signals of H_3O^+ (19+), $H_3O^+.H_2O$ (37+), $H_3O^+.(H_2O)_2$ (55+) and $H_3O^+.(H_2O)_3$ (73+) and dividing the sum by $H_3O^+.(H_2O)_2$ (55+).

3.3. Data sets

The data matrix **X** for all variables, i.e. the combined spectra from 3 precursor ions, contains *n*= 55 samples (rows) and *p*= 701 variables (columns) after removing the hydrated reagent ions. These latter variables were for precursor ion H_3O^+ *m/z* 37 and 55; for O_2^+ *m/z* 32, 37 and 55 and for NO^+ *m/z* 30 and 48. Additionally, three other **X** matrices are created, each consisting of the spectra using one of the three precursors. This allows evaluating the use of individual precursors for their ability to provide spectra discriminating between the different classes. For the H_3O^+ and NO^+ spectra, the **X** matrix contains *n*= 55 samples and *p*= 234 variables. For the O_2^+ spectra, **X** consist of n= 55 samples and *p*= 233 variables. For classification, the **y** vector indicates the three classes, i.e. healthy, cystic fibrosis and asthmatic. Important to notice is that the "raw data" **X** matrix was already normalized by the Syft Technologies proprietary algorithm before other data pretreatment methods were applied. This normalisation includes for every individual ion channel a correction based on a linear quantitative signal, considering both reagent and product ion, as a function of lens voltage, temperature and molecular weight [24].

3.4. Data analysis

Computations were performed with Matlab™ R2014a (The Mathworks, Natrick, MA). All data (pre)processing methods were performed making use of the ChemoAc toolbox 4.1. Modelling of PCA-LDA, PCA-QDA, KNN, CART, PLS-DA and SIMCA, were performed using the classification toolbox 4.2.

4. Results and discussion

Characteristics of the 55 subjects can be found in Table 2 and in Bannier et al. [44], where the same samples were studied by means of an electronic nose.

453 The combined full scan spectra (708 variables) for the 55 samples, belonging to the 3
454 classes, are shown in Figure 2. The variable numbers 1-236 originate from using H_3O^+ as
455 precursor ion, 237-472 from NO^+, while 473-708 were from applying O_2^+.

456 Controlling the humidity is important because differences can lead to varying secondary
457 product ions. Specific secondary product ions (water cluster) for a given VOC could be
458 additionally useful to annotate a given compound. For instance, *m/z* 77 of NO^+ is known
459 as a major water cluster of propanol [45]. However, the ratio of signal levels between an
460 adduct ion and a monohydrate ion depends on the water vapour concentration [46],
461 demonstrating the importance of controlling sample humidity. The humidity could, for
462 instance, be controlled by analysing the samples under two conditions, dry air and moist
463 air (containing a certain percentage of water vapour) [45]. Here, the water concentrations
464 in the samples were measured as an internal control of the analysis [10], occasionally
465 showing any sample introduction issues.

466 As mentioned in the introduction, the diversity between the subjects and their medical
467 treatments with different medicines in combination with other external influences may
468 challenge the classification. The goal of this study is to evaluate which preprocessing and
469 classification techniques are suitable and seems promising to cope with the diversity in the
470 data set and will result in a proper pattern-based classification, useful to implement full-
471 scan SIFT-MS analyses as a diagnostic tool.

472 **4.1. Unsupervised classification**

473 First, the raw data is visualized by PCA to evaluate whether the 3 groups can be
474 distinguished. In Figure 3A, the first PC represents almost 90 % of the total variance. In
475 the PC1-PC2 score plot, the three groups cannot be differentiated. Only two groups are
476 observed along PC1, containing samples of all classes. Two samples, 3 and 8 (asthmatic
477 patients), were separated along PC2 from the two main clusters. This is also seen when
478 only the H_3O^+ spectrum is used. The two deviating samples were not observed in the plots
479 based on the other precursors, while still two groups were present. An explanation for
480 these samples, may be found with the variable 49 (Figure 3B). The precursor and *m/z* of
481 all variables can be found in Supplementary material. Variable 49 has an *m/z* value 65, is
482 formed during the reaction with H_3O^+, as a precursor ion, and does not occur in the other
483 samples.

484 Figure 3B shows the PC1-PC2 loading plot. Two variables were distinct from the rest.
485 Variable 49 seems important for the 2 deviating samples. The identity of this m/z value
486 might be related to methanol or ethanol. The second variable is 484, has *m/z* 30^+ (in
487 O_2^+spectrum) which corresponds to a well-known marker for asthma (nitric oxide) [47,
488 48]. The variable could be discriminant on PC1 for the two clusters observed. However,

unfortunately the observed groups are not related to the classes of interest and the variable is not discriminative for asthmatic patients here.

Note also that SIFT-MS is not a preferred technique for biomarker identification because of the lack of additional annotation purposes, such as a retention time and/or specific mass spectral fragmentation patterns. This results in the possibility that one product ion may be linked to an exhaustive number of compounds. Interesting product ions occasionally may later be identified as potential markers with another technique, such as GC-MS [13]. Furthermore, it is important to understand the ion chemistry in SIFT-MS to know which product ions are related to certain metabolites, even to link to known biomarkers. These biomarkers are not often related to only one product ion in the SIFT-MS spectra [49]. As a result of this uncertainty, only breath metabolites confirmed in other studies are occasionally included as a reference in this pilot study.

The different pretreatments were performed on the raw data with the goal to get an improved classification. After data transformation, the corresponding score plot was visually evaluated. Different methods, as specified higher, were applied. Most did not show the expected groups in the PC1-PC2 score plots. Results were similar as in Figure 3. Some pretreatments, e.g. normalisation, SNV and pretreatment approaches numbers 12-19 (Section 2.1) even resulted in only one observed group in the PC1-PC2 score plot.
A clear distinction between the different classes was found with the DOSC approaches, and Dong's Algorithm followed by DOSC. The O_2^+-precursor-ion spectra (Figure 4) resulted in a score plot where PC1 explained more variation (94 %) than from the spectra based on H_3O^+(77 %), NO^+(86 %) or the combined spectra (86 %). However, all score plots distinguished the three classes. In Figure 4, distinction between the classes is seen along PC1. Samples 3 and 8, which were outlying in the raw data plot (Figure 3) are not outlying anymore along PC1, which determines the class differences, but they increase the variability of the asthmatic group. When determining potential biomarkers, which is not the goal of this study, those responsible for the distinction along PC1, should be searched for, not those increasing the variability along PC2. Again, variable 484, which was discussed higher for the raw data (Figure 3) seems discriminative along PC1, which now distinguishes the three classes. The reason why it is discriminative here and not for the raw data is unclear to us.

The score and loading plots using DOSC after Dong's Algorithm as pretreatment are shown in Figure 5. This preprocessing allowed also visualizing three separated classes along PC1. Figure 5A shows again that samples 3 and 8 are enlarging the asthmatic cluster variability.

The two best pretreatments found were DOSC with and without prior application of Dong's Algorithm. A suitable unsupervised classification may improve the predictive ability of a

classification model, while simpler models can be built [50]. However, a known drawback of DOSC is overfitting [29, 51], which may lead to a perfect class grouping even for randomly assigned classes. We tested the latter for our data set and unfortunately DOSC has led also here to perfect class distinction for randomly assigned classes. This makes the application of DOSC suspicious and dangerous. An explanation for the observation may be that the algorithm searches for data correlated to the classes and that the approach is able to find random correlations which allow distinguishing the randomly assigned classes. For a more thorough evaluation, the supervised classification models, were also built for the randomly assigned classes when DOSC was applied as pretreatment.

4.2. Supervised discrimination and classification techniques

Different classification techniques were applied on the pretreated matrices. First, KNN was considered. Good results were obtained for the matrices pretreated with DOSC, with and without application of Dong's Algorithm (see Table 3). The model based on the O_2^+ spectra classifies all samples correctly both for calibration and cross-validation data. The three other models, certainly the one resulting from the combined spectra, also show good results. However, considering the problems observed with DOSC pretreatments in PCA, the results obtained are suspicious and need further examination (see further).

Another classification technique evaluated is PCA-LDA. The results for DOSC after Dong's Algorithm as pretreatment are also shown in Table 3. For all matrices, except for the NO^+-based, a perfect classification was established. All samples both from the classification set as in cross-validation were correctly classified. Only DOSC as pretreatment resulted in similar results (see in Table 3), The specificity, sensitivity and model efficiency values are 100% for all models (also the NO^+-based). Other preprocessing methods did not result in good PCA-LDA models.

Consecutively, CART and PCA-QDA models were built. A similar output was seen for DOSC pretreatment, with and without application of Dong's Algorithm. The CART models provided good results, with 100% model efficiency for calibration and cross-validation, for the combined and O_2^+ spectra. The other preprocessing methods did not lead to comparable results. The model efficiencies for calibration where only below or around 50%.

PCA-QDA results for the DOSC pretreatments again in perfect predictive classifications. Two other pretreatment methods, approaches 12 and 14, lead to PCA-QDA calibration model efficiencies of 89% for the H_3O^+ matrix. However, the cross-validation efficiencies where only 59%. The results are shown in Table 4 parts A and B. Somewhat better cross-validation results were obtained for the H_3O^+ matrix with Pareto scaling as data pretreatment. Here, 70% model efficiency was obtained for prediction (34 of 55 samples correctly predicted) (Table 4 part C), while calibration showed 82% efficiency (41/55).

The SIMCA and PLS-DA models showed less good predictive abilities with the DOSC pretreatments. Many samples were not classified (more than 50%). These two classification techniques define a threshold for each class [19], as already mentioned in Sections 2.3.4 and 2.3.5. Therefore, samples might be assigned to either none, one or more classes. The first and last situation results in not-assigned samples.

Better PLS-DA calibration results were obtained using other data pretreatments. Model efficiencies of 100% (calibration) were seen for the combined and the H_3O^+ spectra when pretreated with the pretreatment approaches number 14 and 15 as pretreatment. Here, only a limited number of samples is not assigned to a class (see Table 5). However, a concern for these models is the bad results for cross-validation. Similar observations were seen for the approaches number 12 and 13 (Table 6).

Peak annotation based on Variable Importance in Projection (VIP)-scores for the PLS-DA results learned, as already stated, that SIFT-MS is not a suitable technique for biomarker discovery because of the lack of proper peak annotation information. Approximately 40% of the *m/z* values show a VIP score above 1 and would therefore be considered important to distinguish the classes of interest. Consequently SIFT-MS can be applied as phenotyping tool in untargeted full-scan mode or as targeted tool for known compound quantification, but not for biomarker identification.

4.3. Discussion and further evaluation

Because of the possible unreliable results after DOSC preprocessing, further investigation of the applicability of the resulting models as a diagnostic tool is necessary. Further evaluation of the PCA-QDA and PLS-DA models on the Pareto scaled and the combined pretreatments including single value imputation (approaches 12-15) is also needed.

As the best models obtained after DOSC pretreatments are suspicious, they were further evaluated. This pretreatment is known to remove all unwanted variation that is not orthogonal to the class information. As already discussed in the unsupervised section also random classes were perfectly distinguished in the score plot. Additionally, the classification results (for both calibration and cross-validation) of these random classes were similar to those obtained for the real classes, which seemed too optimistic model efficiencies, also in comparison to other diagnostic tools in the literature. Investigation of correlation coefficients learned that the correlation between spectra within a class and between classes were already high. DOSC pretreatment did not lead to an increase of the correlation between samples belonging to the same class nor a decrease between classes, as was observed from color maps. The classes could not be distinguished in these plots while it was expected it would be possible. Dividing the already small data set in a calibration- (41 samples) and test set (14 samples) resulted in similar model performances

for the predictions of the real- and random class models. Here, we had hoped, even though chance correlation is found for the model building when using random classes, that prediction of external validation samples would be worse than for models based on the real classes. Unfortunately, this was not the case for our data. Therefore, the suitability of the DOSC preprocessing technique for the desired classification is not without severe doubt and seems unreliable at the moment. Further research related to the understanding, consideration and applicability of DOSC pretreatment, performed on a large representative data set is thus required.

Consequently, in future, the actual pilot study results should be further investigated applying an extensive data set with enough samples for each class (300 in total). This new data set will allow a proper splitting in representative calibration and test set. It may thus be suitable to further reveal the insights in DOSC pretreatment and allow confirming whether or not it can be used in the investigated classification. This study will also allow to further examine the other approaches, which led to good calibration results but worse predictive ones, on their suitability in this context.

5. Conclusions and future perspectives

SIFT-MS was already used by different research groups as a diagnostic tool for asthma and cystic fibrosis by monitoring specific target compounds. These known breath compounds are nitric oxide, acetic acid, ethanol, methanol, acetone, ammonia, dimethyl disulfide and propanol.

The difference with these targeted studies is that in our actual study the usefulness of SIFT-MS full scan spectra is investigated to discriminate asthma and cystic fibrosis samples from healthy ones. Different data preprocessing techniques in combination with classification techniques are evaluated. The goal was to find a suitable preprocessing method and pattern-based classification model for exhaled-breath diagnosis by SIFT-MS. The knowledge gathered may be of interest to the wider scientific community because data pretreatment and finding good modelling techniques is a labor intensive work.

A possibly interesting data analysis strategy includes building a model (by for instance KNN, CART, PCA-LDA and PCA-QDA) after DOSC pretreatment. Perfect predictive results were found for both calibration- and cross-validation samples. However, the DOSC pretreatment technique has some limitations and led to suspicious results since it allowed also a perfect discrimination of random classes, both in unsupervised analysis and classification modelling. Therefore, other preprocessing techniques were also considered, but they provided less good predictive models by cross-validation. This is for example, the case for PCA-QDA models after using Pareto scaling or the combined preprocessing methods obtaining single value imputation (pretreatment approaches 12 and 14).

Other potentially interesting models are based on PLS-DA after the combined preprocessing methods obtaining single value imputation (pretreatment approaches 12-15). The calibration results were similar to those observed after DOSC pretreatments (100% correct classification), but their cross-validation results were less good.

A future requirement is the collection of an extended data set (about 300 samples), allowing a proper external validation, as well as a further investigation of the results found in the actual pilot study. This larger data set will also allow examining whether the DOSC pretreatment is either perfect or useless as pretreatment for this kind of data.

Nevertheless, this feasibility study showed some potential for the untargeted application of SIFT-MS spectra as rapid pattern-recognition tool, useful in the diagnosis of breath samples.

Acknowledgments

The authors thank all children and parents who give their informed consent to use their breath samples for this study. This work was supported by the Research Foundation Flanders (FWO) under Grant n° G033816N.

Conflict of interest

The authors declare no conflict of interest.

6. References

[1] K.D. Van de Kant, E.M.M. Klaassen, Q. Jöbsis, A.J. Nijhuis, O.C.P. van Schayck, E. Dompeling, Early diagnosis of asthma in young children by using non-invasive biomarkers of airway inflammation and early lung function measurements: study protocol of a case-control study, BMC Public Health, 9 (2009) 1-12. https://doi.org/10.1186/1471-2458-9-210

[2] M. Bannier, K.G. van de Kant, Q. Jöbsis, E. Dompeling, Biomarkers to predict asthma in wheezing preschool children, Clin. Exp. Allergy, 45 (2015) 1040-1050. https://doi.org/10.1111/cea.12460

[3] A.P. Anton, M.D. Sanchez, A.P.C. Pozas, J.L.P. Pavon, B.M. Cordero, Headspace-programmed temperature vaporizer-mass spectrometry and pattern recognition techniques for the analysis of volatiles in saliva samples, Talanta, 160 (2016) 21-27. https://doi.org/10.1016/j.talanta.2016.06.

[4] M.H. Wang, K.C. Chong, M. Storer, J.W. Pickering, Z.H. Endre, S.Y.F. Lau, C. Kwok, M. Lai, H.Y. Chung, B.C.Y. Zee, Use of a least absolute shrinkage and selection operator (LASSO) model to selected ion flow tube mass spectrometry (SIFT-MS) analysis of exhaled breath to predict the efficacy of dialysis: a pilot study, J. Breath Res., 10 (2016), 1-7. http://dx.doi.org/10.1088/1752-7155/10/4/046004

[5] P. Spanel, D. Smith, Progress in SIFT-MS: breath analysis and other applications, Mass Spectrom. Rev., 30 (2011) 236-267. https://doi.org/10.1002/mas.20303

[6] K.H. Kim, S.A. Jahan, E. Kabir, A review of breath analysis for diagnosis of human health, Trac-Trend. Anal. Chem., 33 (2012) 1-8. https://doi.org/10.1016/j.trac.2011.09.013

[7] F. Monedeiro, M. Milanowski, I.-A. Ratiu, H. Zmysłowski, T. Ligor, B. Buszewski, VOC Profiles of Saliva in Assessment of Halitosis and Submandibular Abscesses Using HS-SPME-GC/MS Technique, Molecules, 24 (2019) 2977. https://doi.org/10.3390/molecules24162977

[8] H. Shigeyama, T. Wang, M. Ichinose, T. Ansai, S.-W. Lee, Identification of volatile metabolites in human saliva from patients with oral squamous cell carcinoma via zeolite-based thin-film microextraction coupled with GC–MS, J. Chromatogr. B, 1104 (2019) 49-58. https://doi.org/10.1016/j.jchromb.2018.11.002

[9] J.M. Sanchez, R.D. Sacks, GC analysis of human breath with a series-coupled column ensemble and a multibed sorption trap, Anal. Chem., 75 (2003) 2231-2236. https://doi.org/10.1021/ac020725g

[10] F.J. Gilchrist, C. Razavi, A.K. Webb, A.M. Jones, P. Španěl, D. Smith, W. Lenney, An investigation of suitable bag materials for the collection and storage of breath samples containing hydrogen cyanide, J. Breath Res. 6(3) (2012) 036004. https://doi.org/10.1088/1752-7155/6/3/036004

[11] A.Z. Berna, B. DeBosch, J. Stoll, A.R. Odom John, Breath Collection from Children for Disease Biomarker Discovery, J Vis Exp (144) (2019) e59217. https://doi.org/10.3791/59217.

[12] R. Rodríguez-Pérez, R. Cortés, A. Guamán, A. Pardo, Y. Torralba, F. Gómez, J. Roca, J.A. Barberà, M. Cascante, S. Marco, Instrumental drift removal in GC-MS data for breath analysis: the short-term and long-term temporal validation of putative biomarkers for COPD, J Breath Res 12(3) (2018) 036007. https://doi.org/10.1088/1752-7163/aaa492.

[13] D. Smith, P. Spanel, Selected ion flow tube mass spectrometry (SIFT-MS) for on-line trace gas analysis, Mass Spectrom. Rev., 24 (2005) 661-700. https://doi.org/10.1002/mas.20033

[14] P. Paredi, S.A. Kharitonov, S. Meah, P.J. Barnes, O.S. Usmani, A Novel Approach to Partition Central and Peripheral Airway Nitric Oxide, Chest, 145 (2014) 113-119. https://doi.org/10.1378/chest.13-0843

[15] P. Čáp, K. Dryahina, F. Pehal, P. Španěl, Selected ion flow tube mass spectrometry of exhaled breath condensate headspace, Rapid Commun. in Mass Spectrom., 22 (2008) 2844-2850. https://doi.org/10.1002/rcm.3685

[16] D. Smith, P. Spanel, SIFT-MS and FA-MS methods for ambient gas phase analysis: developments and applications in the UK, Analyst 140(8) (2015) 2573-2591. https://doi.org/10.1039/c4an02049a.

[17] E. van Mastrigt, A. Reyes-Reyes, K. Brand, N. Bhattacharya, H.P. Urbach, A.P. Stubbs, J.C. de Jongste, M.W. Pijnenburg, Exhaled breath profiling using broadband quantum cascade laser-based

spectroscopy in healthy children and children with asthma and cystic fibrosis, J. Breath. Res., 10 (2016) 026003. https://doi.org/10.1088/1752-7155/10/2/026003
[18] M. Kharbach, R. Kamal, M.A. Mansouri, I. Marmouzi, J. Viaene, Y. Cherrah, K. Alaoui, J. Vercammen, A. Bouklouze, Y. Vander Heyden, Selected-ion flow-tube mass-spectrometry (SIFT-MS) fingerprinting versus chemical profiling for geographic traceability of Moroccan Argan oils, Food Chem., 263 (2018) 8-17. https://doi.org/10.1016/j.foodchem.2018.04.059
[19] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Anal. Methods-UK, 5 (2013) 3790-3798. https://doi.org/10.1039/C3AY40582F
[20] P. Oliveri, G. Downey, Multivariate class modeling for the verification of food-authenticity claims, Trac-Trend. Anal. Chem., 35 (2012) 74-86. https://doi.org/10.1016/j.trac.2012.02.005
[21] W. Wu, Y. Mallet, B. Walczak, W. Penninckx, D.L. Massart, S. Heuerding, F. Erni, Comparison of regularized discriminant analysis, linear discriminant analysis and quadratic discriminant analysis, applied to NIR data, Anal. Chim. Acta, 329 (1996) 257-265. https://doi.org/10.1016/0003-2670(96)00142-0
[22] J. Viaene, M. Goodarzi, B. Dejaegher, C. Tistaert, A.H. Le Tuan, N.N. Hoai, M.Chau Van, J. Quetin-Leclercq, Y. Vander Heyden, Discrimination and classification techniques applied on Mallotus and Phyllanthus high performance liquid chromatography fingerprints, Anal. Chim. Acta, 877 (2015) 41-50. https://doi.org/10.1016/j.aca.2015.02.034
[23] M. Bylesjö, O. Cloarec, M. Rantalainen, Normalization and Closure, in: Comprehensive Chemometrics, eds. R. Tauler, B. Walckzak, S.D. Brown, Elsevier, Amsterdam, 2009, pp. 109- 127.
[24] P.-H. Stefanuto, D. Zanella, J. Vercammen, M. Henket, F. Schleich, R. Louis, J.-F. Focant, Multimodal combination of GC× GC-HRTOFMS and SIFT-MS for asthma phenotyping using exhaled breath, Sci. Rep. 10(1) (2020) 1-11. https://doi.org/10.1038/s41598-020-73408-2
[25] S. Wold, M. Sjostrom, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab., 58 (2001) 109-130. https://doi.org/10.1016/S0169-7439(01)00155-1
[26] M. Zeaiter, D. Rutledge, Preprocessing Methods, in: Comprehensive Chemometrics, eds. S.D. Brown, R. Tauler, B. Walczak, Elsevier, Amsterdam, 2009, pp. 121-231.
[27] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers- Verbeke, Handbook of Chemometrics and Qualimetrics, part B, Elsevier, Amsterdam, 1998.
[28] E.M. Kasprzak, K.E. Lewis, Pareto analysis in multiobjective optimization using the collinearity theorem and scaling method, Struct. Multidisc. Optim. 22(3) (2001) 208-218.
[29] J.A. Westerhuis, S. de Jong, A.K. Smilde, Direct orthogonal signal correction, Chemometr. Intell. Lab. , 56 (2001) 13-25. https://doi.org/10.1016/S0169-7439(01)00102-2
[30] T. Galeano-Diaz, M.I. Acedo-Valenzuela, N. Mora-Diez, A. Silva-Rodriguez, Comparative study of different approaches to the determination of robustness for a sensitive-stacking capillary electrophoresis method. Estimation of system suitability test limits from the robustness test, Anal. Bioanal. Chem., 389 (2007) 541-553. https://doi.org/10.1007/s00216-007-1446-1
[31] G. Chen, X. Dong, From chaos to order—perspectives and methodologies in controlling chaotic nonlinear dynamical systems, Int. J. Bifurcat. Chaos, 3 (1993) 1363-1409. https://doi.org/10.1142/S0218127493001112
[32] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers- Verbeke, Handbook of Chemometrics and Qualimetrics, Part A, Elsevier, Amsterdam, 1997.
[33] B.K. Lavine, W.S. Rayens, Classification: Basic Concepts, in: Comprehensive Chemometrics, eds. S.D. Brown, R. Tauler, B. Walczak, Elsevier, Amsterdam, 2009, pp. 507-515.
[34] B.X. Li, Y.H. Wei, H.G. Duan, L.L. Xi, X.N. Wu, Discrimination of the geographical origin of *Codonopsis pilosula* using near infrared diffuse reflection spectroscopy coupled with random forests and k-nearest neighbor methods, Vib. Spectrosc., 62 (2012) 17-22. https://doi.org/10.1016/j.vibspec.2012.05.001
[35] J.S. Shah, S.N. Rai, A.P. DeFilippis, B.G. Hill, A. Bhatnagar, G.N. Brock, Distribution based nearest neighbor imputation for truncated high dimensional data with applications to pre-clinical and clinical metabolomics studies, BMC Bioinformatics, 18 (2017) 1-13. https://doi.org/10.1186/s12859-017-1547-6

753 [36] R. Zdunek, M. Nowak, E. Plinski, Statistical classification of soft solder alloys by laser-induced
754 breakdown spectroscopy: review of methods, J. Eur. Opt. Soci.-Rapid, 11 (2016) 1-20.
755 http://dx.doi.org/10.2971/jeos.2016.16006i
756 [37] M.J. Kangas, R.M. Burks, J. Atwater, R.M. Lukowicz, B. Garver, A.E. Holmes, Comparative
757 Chemometric Analysis for Classification of Acids and Bases via a Colorimetric Sensor Array, J.
758 Chemom., 32(2) (2018) e2961. https://doi.org/10.1002/cem.2961
759 [38] C.-M. Ma, W.-S. Yang, B.-W. Cheng, How the parameters of k-nearest neighbor algorithm impact
760 on the best classification accuracy: In case of parkinson dataset, Journal of Applied Sciences, 14
761 (2014) 171-176. https://doi.org/10.3923/jas.2014.171.176
762 [39] M. Goodarzi, W. Saeys, M.C.U. de Araujo, R.K.H. Galvao, Y. Vander Heyden, Binary classification
763 of chalcone derivatives with LDA or KNN based on their antileishmanial activity and molecular
764 descriptors selected using the Successive Projections Algorithm feature-selection technique, Eur. J.
765 Pharm. Sci., 51 (2014) 189-195. https://doi.org/10.1016/j.ejps.2013.09.019
766 [40] S. Henrard, N. Speybroeck, C. Hermans, Classification and regression tree analysis vs.
767 multivariable linear and logistic regression methods as statistical tools for studying haemophilia,
768 Haemophilia, 21 (2015) 715-722. https://doi.org/10.1111/hae.12778
769 [41] B. Dejaegher, L. Dhooghe, M. Goodarzi, S. Apers, L. Pieters, Y. Vander Heyden, Classification
770 models for neocryptolepine derivatives as inhibitors of the beta-haematin formation, Anal. Chim.
771 Acta, 705 (2011) 98-110. https://doi.org/10.1016/j.aca.2011.04.019
772 [42] S.D. Brown, A.J. Myles, Decision tree modeling in classification, in: Comprehensive
773 Chemometrics, eds. S.D. Brown, R. Tauler, B. Walczak, Elsevier, Amsterdam, 2009, pp. 541-569.
774 [43] A. Giacomino, O. Abollino, M. Malandrino, E. Mentasti, The role of chemometrics in single and
775 sequential extraction assays: A Review. Part II. Cluster analysis, multiple linear regression, mixture
776 resolution, experimental design and other techniques, Anal. Chim. Acta, 688 (2011) 122-139.
777 https://doi.org/10.1016/j.aca.2010.12.028
778 [44] M.A.G.E. Bannier, K.D.G. van de Kant, Q. Jöbsis, E. Dompeling, Feasibility and diagnostic accuracy
779 of an electronic nose in children with asthma and cystic fibrosis, J. Breath Res., 13 (2018), 036009.
780 https://doi.org/10.1088/1752-7163/aae158
781 [45] T. Wang, W. Carroll, W. Lenny, P. Boit, D. Smith, The analysis of 1-propanol and 2-propanol in
782 humid air samples using selected ion flow tube mass spectrometry, Rapid Commun. Mass Spectrom.
783 20(2) (2006) 125-130. https://doi.org/10.1002/rcm.2285
784 [46] D. Smith, K. Sovová, K. Dryahina, T. Doušová, P. Dřevínek, P. Španěl, Breath concentration of
785 acetic acid vapour is elevated in patients with cystic fibrosis, J. Breath Res. 10(2) (2016)
786 021002.https://doi.org/ 10.1088/1752-7155/10/2/021002
787 [47] B. Enderby, D. Smith, W. Carroll, W. Lenney, Hydrogen cyanide as a biomarker for Pseudomonas
788 aeruginosa in the breath of children with cystic fibrosis, Pediatr. Pulmonol. 44(2) (2009) 142-147.
789 https://doi.org/ 10.1002/ppul.20963
790 [48] D. Smith, P. Španěl, On the importance of accurate quantification of individual volatile
791 metabolites in exhaled breath, J. Breath Res 11(4) (2017) 047106. https://doi.org/ 10.1088/1752-
792 7163/aa7ab5
793 [49] D. Smith, M.J. McEwan, P. Španěl, Understanding Gas Phase Ion Chemistry Is the Key to Reliable
794 Selected Ion Flow Tube-Mass Spectrometry Analyses, Anal. Chem. 92(19) (2020) 12750-12762.
795 https://doi.org/10.1021/acs.analchem.0c03050.
796 [50] J. Trygg, J. Gabrielsson, T. Lundstedt, Background Estimation, Denoising, and Preprocessing, in:
797 Comprehensive Chemometrics, eds. S.D. Brown, R. Tauler, B. Walczak, Elsevier, Amsterdam, 2009,
798 pp. 1-8.
799 [51] J. Luypaert, S. Heuerding, D.L. Massart, Y. Vander Heyden, Direct orthogonal signal correction as
800 data pretreatment in the classification of clinical lots of creams from near infrared spectroscopy data,
801 Anal. Chim. Acta 582(1) (2007) 181-189. https://doi.org/10.1016/j.aca.2006.09.029

802

7. Figure captions

Figure 1. Schematic illustration of the Selected-Ion Flow-Tube Mass Spectrometer (SIFT-MS).

Figure 2. Combined full scan SIFT-MS spectra of all measured product ions using the precursors H_3O^+, NO^+ and O_2^+, respectively.

Figure 3. A) PC1-PC2 score plot for the entire raw data matrix. The squares are the cystic fibrosis samples, the dots the healthy samples and the stars the asthmatic patients. B) PC1-PC2 loading plot.

Figure 4. A) PC1-PC2 score plot after Direct Orthogonal Signal Correction (DOSC). The plot is based on the spectra with O_2^+ as precursor ion (233 variables). B) Corresponding loading plot. Symbols: see Figure 3.

Figure 5. A) PC1-PC2 score plot for the matrix with combined spectra (157 variables) after Dong's Algorithm preprocessing followed by Direct Orthogonal Signal Correction (DOSC). B) Corresponding loading plot. Symbols: see Figure 3.

8. Tables

Table 1. *Illustration of the meaning of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) in perspective of class A. Their application in different parameters is specified in Section 2.3.*

		Real class	
		Class A	*Class B/Class C*
Predicted class	*Class A*	TP	FP
	Class B/ Class C	FN	TN

824 ***Table 2.*** *Subject characteristics of the asthmatic and cystic fibrosis patients*

Subject characteristics	Astma	Cystic fibrosis
Number (N)	20	13
Age (mean ± SD)	12.7 ± 3.1	14.4 ± 4.2
Sex: male/female (N)	9/11	11/2
Maintenance therapy ICS alone (%)	100%	31%
Maintenance therapy ICS + LABA (%)	80%	23%
ACT score (mean ± SD)	22.6 ± 4.2	-
ACT score < 20 (%)	20%	-
Pancreatic insufficiency (%)		100%
Pseudomonas aeruginosa colonization (%)		38%
Treatment with maintenance antibiotics (%)		54%
FEV_1 > 90% of predicted (%)		46%
FEV_1 < 70% of predicted (%)		23%

825 *ACT: asthma control test (range 5-25; uncontrolled asthma if score <20); ICS: inhaled*
826 *corticosteroids; FEV_1: forced expiratory volume in 1 second; LABA: long-acting β_2-agonists*

827

Table 3. *Classification parameters for K-nearest neighbours (KNN) and principal component analysis-linear discriminant analysis (PCA-LDA). Pretreatment; Direct Orthogonal Signal Correction (DOSC) with and without combination of Dong's Algorithm. Number of neighbours and of latent variables selected is based on cross validation.*

	KNN (DOSC after Dong's Algorithm)				**PCA-LDA (DOSC after Dong's Algorithm)**				**PCA-LDA (DOSC)**			
Parameters	***All spectra***	***H_3O^+-based spectra***	***NO^+-based spectra***	***O_2^+-based spectra***	***All spectra***	***H_3O^+-based spectra***	***NO^+-based spectra***	***O_2^+-based spectra***	***All spectra***	***H_3O^+-based spectra***	***NO^+-based spectra***	***O_2^+-based spectra***
Number of neighbours/latent variables	3	2	5	1	1	4	3	1	1	4	3	1
Parameters from calibration (%)												
Sensitivity	100	98.18	89.09	100	100	100	90.97	100	100	100	100	100
Specificity	100	98.96	95.02	100	100	100	94.94	100	100	100	100	100
Model efficiency	100	98.75	92.08	100	100	100	90.33	100	100	100	100	100
Number of correct classified samples	55/55	54/55	49/55	55/55	55/55	55/55	50/55	55/55	55/55	55/55	55/55	55/55
Parameters from cross validation (%)												
Sensitivity	98.18	96.36	87.27	100	100	100	90.91	100	100	100	100	100
Specificity	98.79	97.58	93.33	100	100	100	94.95	100	100	100	100	100
Model efficiency	98.33	97.26	90.72	100	100	100	90.33	100	100	100	100	100
Number of correct classified samples	54/55	53/55	48/55	55/55	55/55	55/55	50/55	55/55	55/55	55/55	55/55	55/55

Table 4. *Parameters for principal component analysis-quadratic discriminant analyses (PCA-QDA). Pretreatment: (A) single value imputation by median followed by normalisation by the norm, log transformation and autoscaling; (B) single value imputation by mean followed by normalisation by the norm, log transformation and autoscaling, and (C) pareto scaling.*

	PCA-QDA (A)				**PCA-QDA (B)**				**PCA-QDA (C)**			
Parameters	***All spectra***	***H_3O^+-based spectra***	***NO^+-based spectra***	***O_2^+-based spectra***	***All spectra***	***H_3O^+-based spectra***	***NO^+-based spectra***	***O_2^+-based spectra***	***All spectra***	***H_3O^+-based spectra***	***NO^+-based spectra***	***O_2^+-based spectra***
Number of neighbours/latent variables	4	6	6	5	4	6	2	5	5	6	3	2
Parameters from calibration (%)												
Sensitivity	54.54	85.45	69.09	63.64	57.73	85.45	54.54	63.64	61.82	74.54	49.09	45.45
Specificity	78.18	91.12	82.34	78.87	77.14	91.12	81.26	78.73	84.98	90.69	78.27	77.47
Model efficiency	65.12	88.84	74.93	70.62	63.55	88.84	65.01	69.60	71.75	81.86	59.38	55.89
Number of correct classified samples	30/55	47/55	38/55	35/55	29/55	47/55	30/55	35/55	34/55	41/55	27/55	25/55
Parameters from cross validation (%)												
Sensitivity	38.18	50.91	45.45	50.91	40	50.91	34.54	49.09	47.27	61.82	43.64	41.82
Specificity	68.96	71.43	67.96	71.30	69.52	71.43	67.84	71.17	72.77	80.30	72.90	75.32
Model efficiency	51.73	58.72	52.08	55.22	53.05	58.72	48.50	55.16	59.86	70.41	56.13	53.84
Number of correct classified samples	21/55	28/55	25/55	28/55	22/55	28/55	19/55	27/55	26/55	34/55	24/55	23/55

Table 5. *Parameters for partial least squares-discriminant analysis (PLS-DA). Pretreatment: (A) single value imputation by median followed by normalisation by the norm, log transformation and autoscaling; (B) single value imputation by median followed by probabilistic quotient normalisation, log transformation and autoscaling.*

	PLS-DA (A)				**PLS-DA (B)**			
Parameters	***All spectra***	***H_3O^+-based spectra***	***NO^+-based spectra***	***O_2^+-based spectra***	***All spectra***	***H_3O^+-based spectra***	***NO^+-based spectra***	***O_2^+-based spectra***
Number of latent variables	5	5	5	3	4	5	4	3
Parameters from calibration (%)								
Sensitivity	100	100	98.15	89.74	100	100	98	77.78
Specificity	100	100	98.91	95.61	100	100	98.88	88.73
Model efficiency	100	100	98.70	93.12	100	100	98.63	83.29
Not assigned samples	1.82	1.82	1.82	29.09	5.45	1.82	9.09	18.18
Number of correct classified samples	54/55	54/55	53/55	35/55	52/55	54/55	49/55	35/55
Parameters from cross validation (%)								
Sensitivity	58.33	57.14	47.50	47.37	61.11	54.05	47.37	47.50
Specificity	78.98	72.22	71.83	73.70	79.15	73.73	71.02	73.14
Model efficiency	69.26	63.83	59.73	57.53	71.01	67.26	57.14	57.08
Not assigned samples	34.54	36.36	27.27	30.91	34.54	32.73	30.91	27.27
Number of correct classified samples	21/55	20/55	19/55	18/55	22/55	20/55	18/55	19/55

Table 6. *Parameters for partial least squares-discriminant analysis (PLS-DA). Pretreatment: (A) single value imputation by mean followed by normalisation by the norm, log transformation and autoscaling; (B) single value imputation by mean followed by probabilistic quotient normalisation, log transformation and autoscaling.*

	PLS-DA (A)				**PLS-DA (B)**			
Parameters	***All spectra***	***H_3O^+-based spectra***	***NO^+-based spectra***	***O_2^+-based spectra***	***All spectra***	***H_3O^+-based spectra***	***NO^+-based spectra***	***O_2^+-based spectra***
Number of latent variables	3	5	5	5	4	3	5	4
Parameters from calibration (%)								
Sensitivity	95.92	100	98.15	98.18	100	91.11	98.04	93.88
Specificity	98.68	100	98.91	99.44	100	94.81	98.93	96.33
Model efficiency	97.06	100	98.70	98.76	100	93.55	98.69	94.25
Not assigned samples	10.91	0	1.8	0	5.45	18.18	7.27	10.91
Number of correct classified samples	47/55	55/55	53/55	54/55	52/55	41/55	50/55	46/55
Parameters from cross validation (%)								
Sensitivity	40.54	60	48.72	48.57	58.33	46.15	58.97	45.94
Specificity	69.46	73.36	72.54	79.66	77.67	71.06	77.86	75.15
Model efficiency	51.37	66.73	60.15	59.22	68.99	58.71	64.48	58.68
Not assigned samples	32.73	36.36	29.09	36.36	34.54	29.09	29.09	32.73
Number of correct classified samples	15/55	21/55	19/55	17/55	21/55	18/55	23/55	17/55