Available online at www.sciencedirect.com

ScienceDirect

Journal of Chromatography B, 861 (2008) 101–107

JOURNAL OF CHROMATOGRAPHY B

www.elsevier.com/locate/chromb

Development of accurate classification method based on the analysis of volatile organic compounds from human exhaled air

J.J.B.N. Van Berkel [a], J.W. Dallinga [a], G.M. Möller [b,1], R.W.L. Godschalk [a], E. Moonen [a], E.F.M. Wouters [b,c], F.J. Van Schooten [a,*]

[a] *Department of Health Risk Analysis and Toxicology, Research Institute NUTRIM, University of Maastricht, P.O. Box 616, 6200 MD Maastricht, The Netherlands*

[b] *Department of Respiratory Medicine, University Hospital Maastricht, P.O. Box 5800, 6202 AZ Maastricht, The Netherlands*

[c] *Centre for Integrated Rehabilitation and Organ Failure (CIRO), Horn, The Netherlands*

Received 13 August 2007; accepted 7 November 2007
Available online 19 November 2007

Abstract

Analysis of exhaled air leads to the development of fast accurate and non-invasive diagnostics. A comprehensive analysis of the entire range of volatile organic compounds (VOCs) in exhaled air samples will enable the identification of VOCs unique for certain patient groups. This study demonstrates proof of principle of our developed method tested on a smoking/non-smoking study population. Thermal desorption and gas chromatography coupled to time-of-flight mass spectrometry were used to analyse exhaled air samples. The VOC profiles obtained from each individual were combined into one final database based on similarity of mass spectra and retention indexes (RI), which offers the possibility for a reliable selection of compounds of interest. As proof of principle we correctly classified all subjects from population of smoking ($N = 11$) and non-smoking ($N = 11$) based on the VOC profiles available in their exhaled air. Support vector machine (SVM) analysis identified 4 VOCs as biomarkers of recent exposure to cigarette smoke: 2,5-dimethyl hexane, dodecane, 2,5-dimethylfuran and 2-methylfuran. This approach contributes to future development of fast, accurate and non-invasive diagnostics of inflammatory diseases including pulmonary diseases.

Keywords: VOC; Exhaled air; Breath; Classification; GC/MS; Smokers

1. Introduction

Medical diagnostics and monitoring devices are developing at a fast pace greatly improving public health. Non-invasive analytic methods based on the presence of hundreds of volatile organic compounds (VOCs) in exhaled air could further expand the use of diagnostics. Exhaled air is easily obtained from patients which facilitates repeated sampling of not only the same patient but also larger populations of patients at a lower cost. Many hundreds of VOCs are present in human breath and the opinion is rising that these compounds contain valuable information on an individual's disease status [1]. The presence of some of these VOCs in human breath is thought to be due to degradation of polyunsaturated fatty acids by oxidative stress. This process called lipid peroxidation is a chain reaction process in which reactive oxygen species (ROS) remove an allylic hydrogen atom from lipid membrane structures. This gives rise to a conjugated radical that is peroxidized by oxygen and this way prolongs the chain reaction. Among the final stable reaction products of this process are saturated hydrocarbons like ethane and pentane. These hydrocarbons enter the blood stream and due to their low solubility in blood they are excreted into breath within minutes after formation. Therefore, they could potentially be used to monitor the process of oxidative stress in tissues [2].

One of the first exhaled air related studies was performed by Pauling et al. [3] who identified over 200 compounds present in human exhaled air. Some of these compounds have been associ-

Abbreviations: VOC, volatile organic compounds; RI, retention index; SVM, support vector machines; ROS, reactive oxygen species; GC–TOF–MS, gas chromatograph–time-of-flight–mass spectrometer; RT, retention time; MF, match factor.

[*] Corresponding author.
E-mail address: F.vanSchooten@GRAT.unimaas.nl (F.J. Van Schooten).

[1] Present address: Department of Respiratory Medicine, Leiden University Medical Center, Leiden, The Netherlands.

1570-0232/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2007.11.008

ated with different pathological conditions. For instance ethane and pentane levels have been linked to oxidative stress and lipid peroxidation [4] and a decrease of exhaled isoprene levels correlated with exacerbations of cystic fibrosis [5]. In 1985, Gordon et al. identified alkanes and monomethylated alkanes in exhaled air of lung cancer patients [6], stating the use of the identified compounds as possible biomarkers. In 1999, Phillips et al. selected 22 VOCs to classify subjects with and without lung cancer [7], and in 2003 modified the VOC pattern by reducing their number to nine [8]. More recently in 2007 Phillips et al. concluded that volatile biomarkers in breath were sensitive and specific for pulmonary tuberculosis [9]. In 2006 Barker et al. proved the feasibility of chemical breath analysis for VOCs as they studied 12 volatile compounds in exhaled air in relation to cystic fibrosis. Only one component demonstrated to be significantly different in CF patients compared to healthy subjects [10]. We developed a more accurate approach of investigating the full range of VOCs in exhaled air and obtained proof of principle by correctly classifying human breath of smokers and non-smokers.

2. Experimental

2.1. Study subjects

A total of 22 subjects, 11 smokers and 11 non-smokers free from chronic lung disease or respiratory tract infection, as confirmed by medical history, were included in this study (Table 1). No restrictions were applied regarding drugs, alcohol or diet. Subjects were all sampled at one centrally ventilated room at the university. Participation to this study was voluntary. The authors are aware of the small group size, but this study is setup to provide analytical proof of principle of the methodology presented here and will in the future be used on very large subject groups.

2.2. Sample collection and analysis

Exhaled air was collected by exhaling into inert Tedlar bags (5 L). Subjects were asked to inhale, hold their breath for 5 s and subsequently fully exhale into the Tedlar bag. All Tedlar bags were washed twice with high-grade nitrogen as described by the manufacturer before usage to make sure all contaminants were eliminated. The content of the Tedlar bag was transported under standardized conditions onto desorption tubes; stainless steel two-bed sorption tubes, filled with carbograph 1TD/Carbopack X (Markes International, Llantrisant, Wales, UK). These desorption tubes were placed inside the thermal desorption unit (Marks Unity desorption unit, Marks International Limited, Llantrisant, Wales, UK) and quickly heated to 270 °C in order to release all VOCs and transport the released VOCs onto the GC-capillary. The used desorption unit was highly suitable for repeated, quantitative and reproducible measurements. Ten percent of the sample was injected into the GC, the remaining 90% transported to another adsorption tube for storage and may be used for later reanalysis. Just before the sample enters the GC the sample is trapped by a cold trap at 5 °C in order to concentrate the sample. Next VOCs were separated by capillary gas chromatography (column: RTX-5ms, 30 m × 0.25 mm 5% diphenyl, 95% dimethylsiloxane, film thickness 1 μm, Thermo Electron Trace GC Ultra, Thermo Electron Corporation, Waltham, USA). The temperature of the gas chromatograph was programmed as follows: 40 °C during 5 min, then raised with 10 °C/min until a final maximum temperature 270 °C in the final step this temperature was maintained for 5 min. Time-of-flight mass spectrometry (TOF-MS) (Thermo Electron Tempus Plus time-of-flight mass spectrometer, Thermo Electron Corporation, Waltham, USA) was used to detect and identify components available in the samples. Electron ionization mode was set at 70 eV and the mass range *m*/*z* 35-350 was measured. Sample frequency of the mass spectrometer was set to 5 Hz and analysis run time to 33 min.

Table 1
Study subject characteristics

	Smoking (*n* = 11)	Non-smoking (*n* = 11)
Age (years)	54 ± 13	47 ± 11
Packyears	26 ± 19	–
Sex (m/f)	4/7	6/5

2.3. Data-acquisition and data mining

Analysis of the data output files from the GC–TOF–MS was performed in successive steps as described below.

2.3.1. Peak detection and corrections

Automated peak detection and baseline correction were performed on the chromatographic raw GC/MS output data files. Baseline correction adjusts the variable background by the following steps: first the background is estimated within multiple shifted windows of width 200 *m*/*z*, next the varying baseline is regressed to the window points using a spline approximation, and finally the background of the input signal is adjusted. Peak detection consisted of first smoothing the signal. After this step peak locations were assigned. Finally peaks not satisfying specific criteria, like full peak width at half height and maximum base width were eliminated.

The raw GC/MS files contain mass spectra at every MS-scan performed (sample frequency of 5 Hz). The resulting output was saved to a file containing detected peak areas and respective scan numbers. To combine mass spectra and areas belonging to the detected peaks the raw GC/MS files and the peak detection output files were merged through a combination of scan numbers. This resulted in a file containing four columns: scan numbers, retention times (RT), peak areas and mass spectra belonging to the detected peaks.

2.3.2. Normalization of retention time within a sample run and between subjects' chromatograms

Normalization of retention times (RT) to retention indices (RI) is necessary to reduce the instrumental variation by adjusting the retention times within each sample run. This was achieved by normalizing RT to the toluene retention time.

Next the data were corrected for chromatographic drifting by determining retention indices of 13 widely available compounds (acetone, 1-propanol, benzene, toluene, furfural, xylene, styrene, heptanal, phenol, D-limonene, decanal, diethylphtalate, diphenylsulfide) in each chromatogram. RI times of these compounds were used in applying corrections in order to line-up the RI indices of all the sample files against one reference sample file. Polynomial functions and interpolation was used to obtain the best fit and correct all RT entries.

2.3.3. Matching peaks based on similarity of mass spectra and retention indices

Subsequently all the corrected files – one for each exhaled air sample – were combined into one large database file by lining up all calculated peak areas of the according compounds based on RI window settings and similarity match factors (MFs) between mass spectra. The MFs between mass spectra were calculated using the best performing routine according to Stein and Scott [11]; the dot-product function that measures the cosine of the angle between spectra were represented as vectors.

In order to combine the output files of all individuals into one working file suited for statistical analysis, one file was chosen as reference file based on the overall quality of the measurement. Next a second output file was selected. Compounds from this second file were to be combined with the complementary compounds from the reference file. Combination of these compounds was based on mass spectra similarity with use of the MF-values and the potentially complementary compounds needed to be within a certain RI-range. If no good fit was found the peak was added to the reference file as a new entry. This data combination routine was repeated for every file to be included. Finally the resulting dataset was checked for RI inconsistencies and compounds demonstrating these RI-inconsistencies were removed if necessary. This RI-inconsistencies-check is based on the fact that if the same instrumental procedure is used for the analysis of different samples the RI-order of the detected compounds must be the same.

2.3.4. Quantification of peak areas

After all the corresponding peak areas of the complementary compounds were combined into one large dataset, normalization of the peak area data was performed in order to be able to compare the different peak areas from different samples. This was necessary because the exhaled air samples contained different unknown absolute volumes of exhaled air, which makes comparison of amounts of compounds impossible. Another reason for normalization is to correct for fluctuations in the response of the mass spectrometer.

Different types of global normalization have been evaluated. The most promising rescaling factor used in this study is based on the cumulative area under the detected peaks and implemented into the final database file. Since all chromatograms display rather similar profiles this method of normalization is most robust. Another benefit regarding this area scaling factor is that it does account for the baseline noise present in the raw chromatographic signal.

A measure to rule out most of the noise resulted in discarding peaks with RI < 0.15 and RI > 2.8. The deleted noise was mostly due to a high degree of column bleeding after RI > 2.8. Also the very light compounds that elute from the column before RI < 0.15 usually contain noisy mass spectra in our setup.

2.3.5. Classification model

To determine which compounds added to the database were of interest with regard to the classification of smoking and nonsmoking subjects, we used support vector machines (SVM). Several experiments have been performed with different classifiers like random forest, discriminant analysis and principal component analysis. These experiments demonstrated SVM to outperform all others regarding compound selection. SVM was able to select those compounds that provided the best performance as implemented into a classifier. SVM demonstrate the ability to construct predictive models with large generalization power even in the case of large dimensionality of the data or when the number of observations available for training is low. SVM always seeks a globally optimized solution and avoid overfitting. This implies a large number of features (i.e. compounds) are allowed. This encouraged us to implement this subset selection algorithm into this study. This algorithm will select the most optimal subset of compounds able to correctly classify our dataset. A variety of subset selection methods was tested, like gain-ratio attribute evaluator. The best subset of compounds was selected using the attribute selection option implemented in Weka [12]: a collection of machine learning algorithms for data mining tasks. Attributes (compounds) were selected using an SVM attribute evaluator. The attribute evaluator we used evaluated the worth of a subset of attributes by considering the individual predictive ability of each feature along with the redundancy between them. Preferably features will be selected showing high correlations within the class and low intercorrelation. Next the selected attributes were analyzed and ranked with use of SVM using recursive feature selection and removing one attribute at a time. This way attributes were selected using the weight magnitude as ranking criterion. After every run the least efficient attribute was removed. All resulting subsets were analyzed for classification performance with use of support vector classifiers based on John Platt's sequential minimal optimization algorithm and the random forest classification algorithm [13].

3. Results

3.1. Reproducibility and variability

To validate the newly developed method to extract the discriminating compounds, the instrumental reproducibility and inter- and intra-individual variability were tested as well differences in exhalation patterns.

3.1.1. Instrumental reproducibility

Instrumental reproducibility was determined by analyzing identical exhaled air samples that were obtained by emptying a filled bag over a y-shaped connector onto two absorption tubes. The two absorption tubes were subsequently analyzed by GC–TOF–MS. This experiment was repeated six times. The instrumental reproducibility was demonstrated by comparing the two complementary chromatograms as demonstrated in

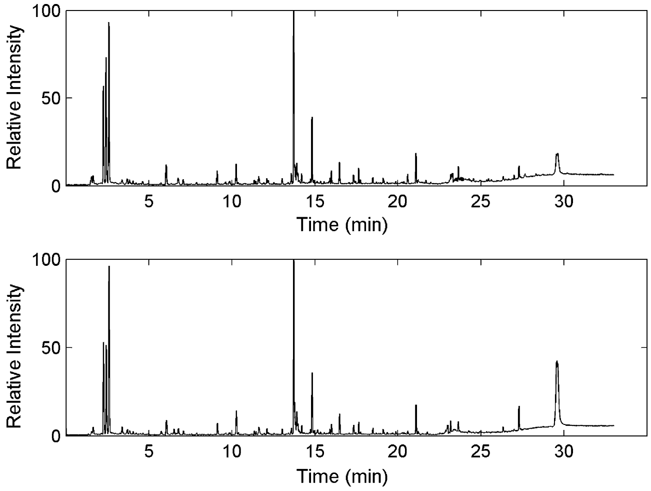

Fig. 1. Example of two chromatograms demonstrating instrumental reproducibility. The measured samples contained identical exhaled air samples. Visual inspection confirms high degree of similarity.

Fig. 1. Already from visual inspection of the two chromatograms it can be concluded that the two chromatograms are highly similar, confirming a high degree of instrumental reproducibility. The quantification of the similarity was done by means of calculation of a distance measure (dot product rule) and is presented in the boxplot of Fig. 4. This distance measure is based on the similarity of the entire raw chromatogram. Distance measure calculation of all complementary files resulted in a distance measure ranging from 0.96 to 0.99. A value of '1' denotes identical samples, the lower the value the lesser the degree of similarity.

3.1.2. Inter- and intra-individual variability in VOC profiles

Intra-individual and inter-individual variability were also mapped. Intra-individual variability was examined by repeated sampling of exhaled air from 10 non-smoking subjects for 5 consecutive days and comparing the results per subject from day to day. Inter-individual variability was examined by sampling 10 non-smoking subjects and comparing the data from subject to subject. Examples of the resulting chromatograms are shown. In Fig. 2a one subject sampled at 2 consecutive days is presented and it can be seen that the two chromatograms show a high degree of similarity. Fig. 2b shows chromatograms from two different subjects sampled in the same room at the same time. Shown chromatograms demonstrate that the degree of similarity is less as compared to the chromatograms of Fig. 2a.

Again the similarity between several chromatograms was quantified using a distance measure as previously mentioned. The results regarding inter-individual and intra-individual variability are shown in Fig. 4. This figure shows boxplots representing (a) instrumental reproducibility match factors, (b) exhalation flow rate depended match factors (c) intra-individual variability match factors and (d) inter-individual variability match factors. As expected it can be seen from this figure that the intra-individual variability ranging from 0.80 to 0.99 is far smaller than the inter-individual variability ranging from 0.16 to 0.98; this is consistent with previously performed studies [14,15].

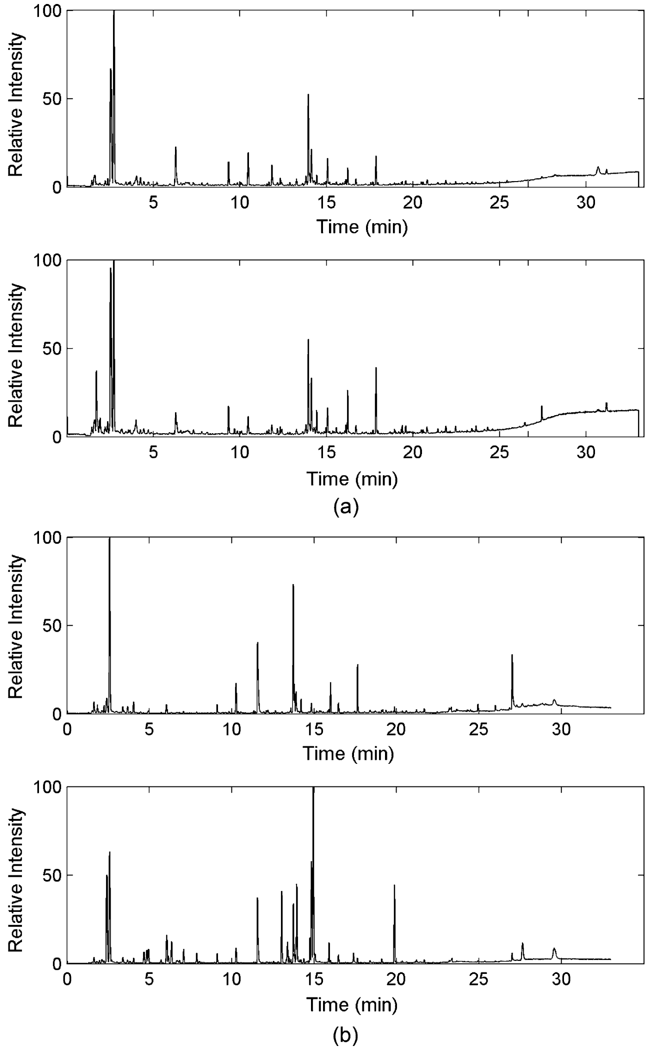

Fig. 2. Examples of representative chromatograms from (a) a subject sampled at 2 consecutive days to examine intra-individual variability and (b) two different subjects sampled at the same time to examine inter-individual variability.

3.1.3. Exhalation characteristics and its impact on VOC profiles

In order to determine whether standardization of the sampling method of the subjects is necessary, an experiment was performed to explore the effect of different exhalation patterns on VOC profiles. To determine whether differences in exhalation air sampling of subjects was a variable in our newly developed methodology 5 non-smoking subjects inflated 2 Tedlar bags as follows: one bag was inflated by superficial exhalation and the other one was inflated after deep inspiration, a 5 s breath hold and subsequent total exhalation into the sample bag as suggested by Barker et al. [10]. This procedure was repeated five times with approximately 90 min intervals in the same centrally ventilated room. In Fig. 3 the resulting chromatograms are shown and as judged already from visual inspection it can be concluded that these complementary chromatograms demonstrate a high degree of similarity suggesting that superficial and deep exhalation are resulting in similar VOC profiles.

Mann–Whitney testing showed that only 58 out of the total of 1201 overall detected compounds proved to be statistically dif-

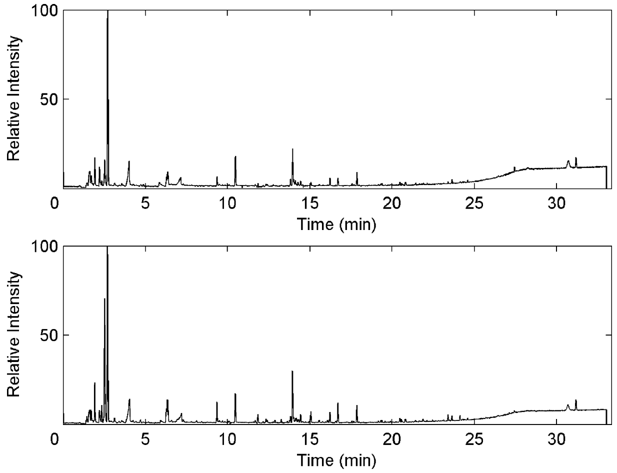

Fig. 3. Examples of representative chromatograms from a subject inflating a bag superficially (upper graph) and deeply (lower graph).

ferent ($p \leq 0.05$) for the two different exhalation methods. After correcting for multiple testing by applying Bonferroni correction for the alpha-value no compound proved to be significantly affected by the exhalation characteristics. Again quantification of the similarity was done by means of calculation of a distance measure and is presented in the boxplot of Fig. 4b. As can be seen the degree of similarity is comparable to the intra-individual similarity, proving difference in exhalation patterns did not lead to significant difference in VOC profiles within an individual.

3.1.4. Validation of methodology on exhaled air from smokers and non-smokers

To validate the methodology we analyzed the exhaled air from 11 smoking and 11 non-smoking subjects. The subjects exhaled a mean of 381 identified different VOCs. All 22 subjects were combined into one large database. In order to correctly combine the peaks from different subjects an MF-threshold value of 0.85 and an RI-window value of 0.045 were used. This resulted in a database of 22 subjects and 3211 compounds, 467 compounds were present in at least 5 of the 22 subjects.

Compounds that were detected in only 2 of the subjects or less were discarded since these compounds do not exert any discriminatory power due to their low occurrence rate and might introduce noise if implemented into the classification model. This value of at least 3 times availability has been introduced by trial and error testing of different threshold values and from similar experiments as mentioned in literature [16]. Applying this threshold criterion resulted in a database consisting of 1095 components. The selection of peaks that discriminate smokers from non-smokers as described in Section 2.3.4 was based on this final database. The most optimal classification model was based on a support vector classifier using just 4 VOCs. This model classifies all subjects correctly regarding their smoking behavior as tested with a 10 times cross validation. Other classification models like random forest, random tree, multilayer perceptrons and Bayesian classifier were also used but the various classifiers tested did not yield improved performances. The same observation was reported by Guyon et al. [17]. Since the model based on SVM outperformed other classifiers, this type of classifier was selected.

We identified VOCs implemented into the classification model with spectrum recognition using the NIST library in combination with spectrum interpretation by an experienced mass-spectrometrist and identification based on retention times of components. Table 2 shows the identified VOCs.

Fig. 5 shows the relative amount of the relevant compounds available in the exhaled air. Bars left from the dotted line represent non-smoking subjects, bars to the right of the dotted line represent the smoking group; the height of the bar represents the normalized integrated peak area of the selected component. As can be seen from Fig. 5 the combined classification power of these four compounds is in most subjects based on availability in

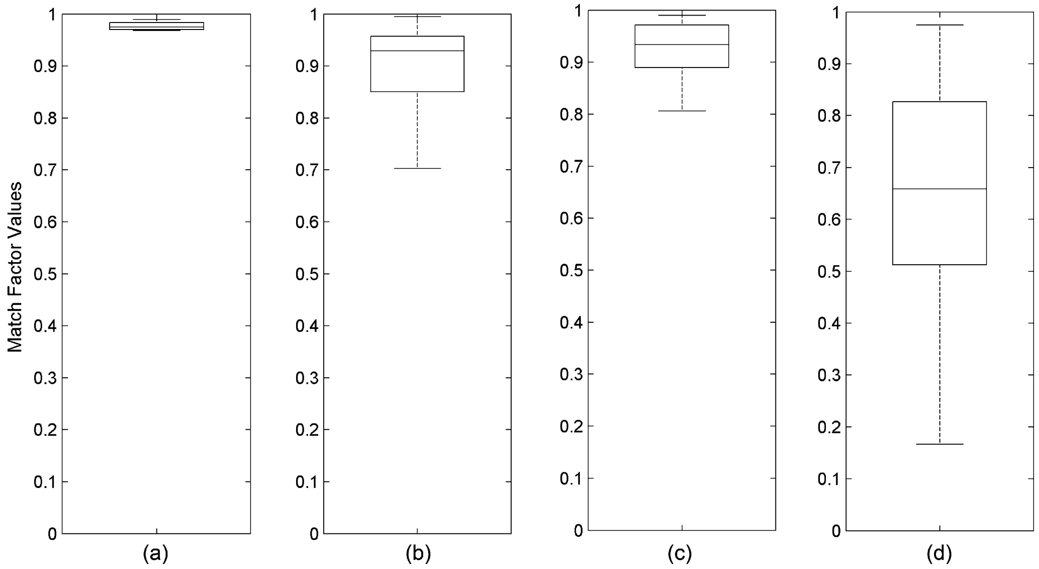

Fig. 4. Boxplots of match factors that are based on similarity between raw chromatograms. Boxplot representing (a) instrumental reproducibility match factors, (b) exhalation flow rate depended match factors, (c) intra-individual variability match factors and (d) inter-individual variability match factors. As demonstrated the instrumental reproducibility (a) is by far the smallest, and as expected the inter-individual variability is larger (d) compared to the intra-individual variability (c).

Table 2
Compounds used in the classification model to classify smokers and non-smokers using VOCs in exhaled air

Compound name	Retention time (min)	No. of times detected in 22 samples
2,5-Dimethyl hexane	7.98	11
Dodecane[a]	17.45	17
2,5-Dimethylfuran	7.46	8
2-Methylfuran	4.16	7

[a] Confirmed by retention index.

smoking subjects versus absence or levels below detection limit of these compounds in exhaled air of non-smoking subjects. As can be seen from Fig. 5 the individual classification power of each compound is not 100%. The SVM implementing and combining data from all 4 compounds is however able to classify all subjects correctly.

4. Discussion

VOCs in exhaled air are thought to represent several processes in the human body, like metabolism and lipid peroxidation, and therefore have a great potential as non-invasive biomarkers of human health, presence and possibly severity of disease. In this paper a newly developed method for analyzing and data processing of exhaled air samples has been presented and tested on a small validation set of exhaled air samples of 11 smoking and 11 non-smoking subjects. We are aware that the analysis of exhaled air and more specifically the analysis of VOCs from exhaled air and relating them with disease is not a new approach [3,7]. We want to emphasize that in the present study, a more robust method was developed for sampling and data mining of the acquired data then was published until now. One of the main advantages of our approach is that raw mass spectra are used to find complementary compounds in all subjects, instead of combining compounds based on identity. The match factor as described by Stein et al. [11] is implemented to determine the degree of similarity between measured mass spectra instead of comparison against library values. We experienced that first identifying the compounds and then finding complementary compounds in all samples based on compound names, introduced more mismatches compared to matching based on the raw mass spectra. We are confident that comparing the retention times and match factors will result in more correctly combined compounds in the respective subjects.

The selected subjects exhaled into a Tedlar bag and volatile organic compounds were trapped on desorption tubes and analyzed with use of a gas-chromatograph in line with a time-of-flight mass spectrometer. The resulting data were processed using newly developed routines. To validate this analytical method several reproducibility and variability measurements were performed to assess instrumental variability and both inter- and intra- individual variability. As demonstrated in Fig. 4 the instrumental variability (a) is very small which confirmed the high reproducibility of our technology. As expected, inter-individual variability is larger than intra-individual variability and both show greater variation than instrumental variability, again confirming the reliability of our methodology.

Other studies detecting VOCs in exhaled air mentioned the necessity to correct for chemical background appearing in their samples. In our case, no background corrections have been taken into account. This is due to the fact that it will not be possible

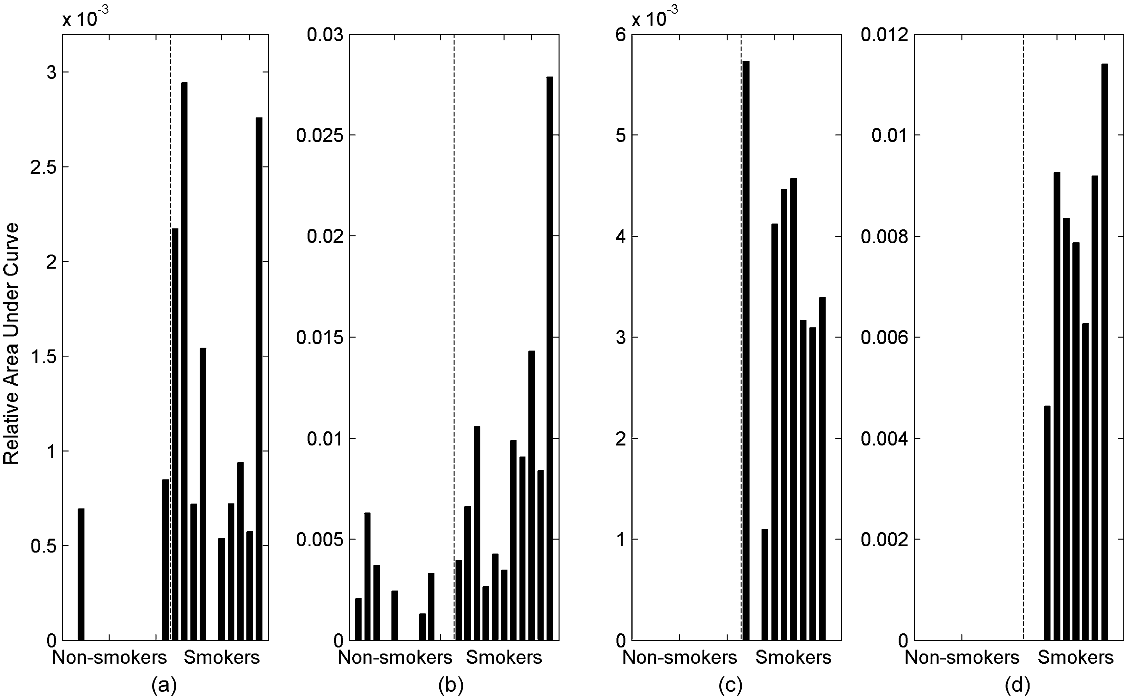

Fig. 5. Compound availability of identified discriminatory compounds used in an SVM model able to classify all smoking/non-smoking subjects correctly. Left sides of the graphs depict the non-smoking subjects, the right sides depict the smoking subjects. (a) 2,5-Dimethylhexane, (b) dodecane, (c) 2,5-dimethylfuran, (d) 2-methylfuran.

to correct for the complex interdependencies between excretion and uptake of VOCs by easily subtracting the inhaled from the exhaled air [2]. Moreover, background noise will be randomly distributed between subjects' samples and would thus neither exert any discriminatory power, nor interfere with the outcome of the analyses. We are aiming with discriminative analysis only to select those compounds that are specific for the disease or condition and should thus principally not depend on background chemicals.

The data-analysis design was finally tested on a dataset containing 1095 VOCs from 11 smoking and 11 non-smoking subjects. After classification analysis, a support vector classifier based on only 4 compounds – identified as 2,5-dimethylhexane, dodecane, 2,5-dimethylfuran and 2-methylfuran – was able to correctly classify all subjects based on 10-times cross validation. The authors are aware that simpler statistical approaches like *T*-statistics or discriminant analysis will perform similar in a small group size as the one used in this study. But since this methodology was designed to be used on large groups with hundreds of subjects a powerful approach like SVM was chosen. We are aware of the fact that use of an SVM classifier to correctly classify 22 subjects is a bit overpowered, but here we merely provide a proof of principle.

The origin of the discriminating compounds in exhaled air remains unclear so far, although these compounds have been identified previously in relation to smoking. In 2002 Gordon et al. already demonstrated 2,5-dimethylfuran to be a promising breath biomarker in detection of active smoking [18] and Sanchez et al. in 2006 identified 2,5-dimethylfuran and 2-methylfuran as strong indicators of smoking status [14]. Although it is well known that active cigarette smoking directly affects the levels of benzene and other VOCs in breath of smokers and previous research demonstrated that concentrations of benzene detected in exhaled air of smokers are always higher than of non-smokers [15,19], benzene and other important constituents of cigarette smoke have not been included in our most optimized model. The exclusion of for example benzene is because this model represents the best subset of compounds that provides the most optimal classification, also taken the redundancy of the compounds into account.

In conclusion, this study demonstrated the functionality of our approach of exhaled air analysis by demonstrating discrimination based on smoking status of subjects. The presented methodology is very accurate and has great power. This design regarding the analysis and identification of discriminatory biomarkers in exhaled air might allow for non-invasive monitoring of inflammation and oxidative stress in the respiratory tract in patients suffering from (inflammatory) lung diseases.

Acknowledgements

The authors acknowledge the province of Limburg for the financial support of this research at the University of Maastricht.

References

[1] W. Cao, Y. Duan, Clin. Chem. 52 (2006) 800.

[2] W. Miekisch, J.K. Schubert, G.F. Noeldge-Schomburg, Clin. Chim. Acta 347 (2004) 25.

[3] L. Pauling, A.B. Robinson, R. Teranishi, P. Cary, Proc. Natl. Acad. Sci. U.S.A. 68 (1971) 2374.

[4] L.T. McGrath, R. Patrick, P. Mallon, L. Dowey, B. Silke, W. Norwood, S. Elborn, Eur. Respir. J. 16 (2000) 1065.

[5] C.M. Kneepkens, C. Ferreira, G. Lepage, C.C. Roy, Clin. Invest. Med. 15 (1992) 163.

[6] S.M. Gordon, J.P. Szidon, B.K. Krotoszynski, R.D. Gibbons, H.J. O'Neill, Clin. Chem. 31 (1985) 1278.

[7] M. Phillips, K. Gleeson, J.M. Hughes, J. Greenberg, R.N. Cataneo, L. Baker, W.P. McVay, Lancet 353 (1999) 1930.

[8] M. Phillips, R.N. Cataneo, A.R. Cummin, A.J. Gagliardi, K. Gleeson, J. Greenberg, R.A. Maxfield, W.N. Rom, Chest 123 (2003) 2115.

[9] M. Phillips, R.N. Cataneo, R. Condos, G.A. Ring Erickson, J. Greenberg, V. La Bombardi, M.I. Munawar, O. Tietje, Tuberculosis (Edinb) 87 (2007) 44.

[10] M. Barker, M. Hengst, J. Schmid, H.J. Buers, B. Mittermaier, D. Klemp, R. Koppmann, Eur. Respir. J. 27 (2006) 929.

[11] S.E. Stein, D.R. Scott, J. Am. Soc. Mass Spectrom. 5 (1994) 859.

[12] E. Frank, M. Hall, L. Trigg, G. Holmes, I.H. Witten, Bioinformatics 20 (2004) 2479.

[13] J.C. Platt, Fast Training of Support Vector Machines using Sequential Minimal Optimization, MIT Press, 1998.

[14] J.M. Sanchez, R.D. Sacks, Anal. Chem. 78 (2006) 3046.

[15] L. Wallace, T. Buckley, E. Pellizzari, S. Gordon, Environ. Health Perspect. 104 (Suppl. 5) (1996) 861.

[16] D.J. Penn, E. Oberzaucher, K. Grammer, G. Fischer, H.A. Soini, D. Wiesler, M.V. Novotny, S.J. Dixon, Y. Xu, R.G. Brereton, J. R. Soc. Interface 4 (2007) 331.

[17] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Mach. Learn. 46 (2002) 389.

[18] S.M. Gordon, L.A. Wallace, M.C. Brinkman, P.J. Callahan, D.V. Kenny, Environ. Health Perspect. 110 (2002) 689.

[19] L. Wallace, Environ. Health Perspect. 104 (Suppl. 6) (1996) 1129.