Journal of Chromatography, 511 (1990) 291–302
Elsevier Science Publishers B.V., Amsterdam

CHROM. 22 442

Identification of exposure markers in smokers' breath

SYDNEY M. GORDON
IIT Research Institute, 10 West 35th Street, Chicago, IL 60616-3799 (U.S.A.)
(Received December 14th, 1989)

ABSTRACT

Volatile organic compounds present in the exhaled breath of 26 smokers and 43 non-smokers were evaluated in an effort to identify possible biochemical markers resulting from the exposure to cigarette smoke. The total ion current profiles obtained from gas chromatography–mass spectrometry (GC–MS), which contained about 230 GC–MS peaks, were first analyzed by using standard statistical procedures to select a subset of 22 peaks. The importance of the peaks was ranked using factor analysis, which further reduced the dimensionality of the data, and discriminant analysis served to develop classification functions. One peak, 2,5-dimethyl furan, had sufficient discriminatory power in the GC–MS profiles to allow almost complete differentiation (96% correct classification) between the smokers and non-smokers groups. In addition, several other compounds were able to separate the groups with a high level of accuracy.

INTRODUCTION

The major adverse effects of cigarette smoking on human health (*e.g.*, lung cancer, impaired lung function) are dose-related[1,2]. Although cessation of smoking substantially reduces the risk of lung cancer among smokers, the benefits of smoking reduction programs have to be deduced from unvalidated questionnaires and are therefore sometimes difficult to assess. Hence, independent quantitative markers are needed to validate survey data and monitor compliance behavior.

Of the nearly 4000 individual components identified in cigarette mainstream smoke[3], the uptake of only a few markers has been studied. Biochemical procedures have been developed to measure the concentrations of nicotine, cotinine, thiocyanate, carboxyhemoglobin and carbon monoxide in blood, saliva, expired air and hair[4–11]. While all these procedures register differences between smokers and non-smokers, they are either invasive, analytically complex or non-specific. The potential of alternative methods and the existence of other markers should therefore be explored.

Human breath is an important mode of uptake and elimination of volatile organic compounds (VOCs) in exposed individuals[12]. Numerous studies have demonstrated that breath analysis provides a powerful means of establishing an

0021-9673/90/$03.50

unequivocal diagnosis of exposure and, in some cases, for estimating the extent of exposure[13]. Analysis of exhaled breath, rather than blood or urine, is attractive for two reasons. First, it is non-invasive and therefore preferable for use in studies relying on reasonable levels of response from volunteers among the general public. Second, since breath is a less complex mixture than other body fluids, it is generally more amenable to comprehensive chemical characterization.

Despite the apparent potential offered by breath analysis for diagnosing exposure, its use to date in the objective measurement of exposure to cigarette smoke has been limited. Most earlier work was based on the assumption that only one or two constituents were relevant[4,14]. However, information on concurrent changes in the concentration of several constituents could potentially increase the diagnostic significance of expired air analysis in distinguishing smokers from non-smokers. In a recent study, personal air exposures and exhaled breath concentrations of 25 VOCs were measured by gas chromatography–mass spectrometry (GC–MS) for 200 smokers and 322 non-smokers in New Jersey and California[15–17]. Smokers showed significantly elevated breath concentrations of benzene, styrene, ethylbenzene, *o*-xylene and *m*- + *p*-xylene. The GC–MS data were generated as part of the U.S. Environmental Protection Agency's TEAM (Total Exposure Assessment Methodology) Study, which was statistically designed to measure exposures and corresponding breath levels of about 600 persons in several U.S.A. cities, representing a total population of 700 000 residents[15]. Although measurements were made on 100–200 volatile chemicals in the breath of each participant, the TEAM study focused on only 25 specific compounds. The results obtained by Wallace and co-workers[15–17] for benzene and the other target compounds suggest that there may be other volatile compounds present in exhaled breath that are even more characteristic of smoking activity.

Because of the volume and complexity of the data, as well as the large variations in data commonly found among individuals in a given sample population, special GC–MS data-enhancement procedures together with multivariate statistics (pattern recognition) are needed to extract useful information from the data files and differentiate between smokers and non-smokers. In an earlier investigation[18], we demonstrated the application of these techniques to high resolution GC–MS breath data from lung cancer patients. The method identified several VOCs in the breath of lung cancer patients that had sufficient diagnostic power in the GC–MS profiles to allow almost complete differentiation between them and a group of controls. Similarly, this study has sought to identify exposure markers in the breath of smokers that may be used to distinguish them from non-smokers in non-invasive exposure screens.

EXPERIMENTAL

As the quality assurance laboratory for the TEAM study, we analyzed roughly 20% of all the breath samples collected during the investigation. It was this subset of the complete data base of measured breath VOCs that we evaluated for discriminants between smokers and non-smokers.

Subject selection and characterization

The subjects were selected by the prime contractor (Research Triangle Institute)

TABLE I

CHARACTERISTICS OF PARTICIPANTS

Category		*Number of respondents*
Smoking status	Smoker	26
	Non-smoker	43
Smoked during monitoring period	Yes	26
	No	43
Exposed to tobacco smoke	Yes	37
	No	32

by a three-stage probability sampling design, and were stratified to provide a geographic dispersion of the residences sampled within each municipality[19–23]. The probability sampling methods they used were designed to ensure that the data collected could be used to draw valid statistical inferences concerning the populations sampled. For the present purposes, therefore, the subset population studied is assumed to represent a suitably randomized sample.

After being interviewed by the prime contractor, all participants completed a questionnaire describing their occupations and kept a record of their activities during the monitoring period[19–23]. The data provided the secondary stratification variables of potential occupational exposure and smoking status, as well as age, sex and race. The questionnaire also gave information, *inter alia*, on the number of cigarettes smoked per day, the age at which smoking first started and, where applicable, the age at which smoking stopped. A 24-h activity screener provided information on the use of tobacco during the preceding 24-h period. Table I presents a summary of relevant characteristics of the participants in the sample subset. Of the total, 38% were current smokers, all of whom had smoked during the monitoring period, while 54% had been exposed to tobacco smoke during the preceding 24-h period.

Sample collection and analysis

Sample collection was carried out by the prime contractor and has been described in detail elsewhere[15,22]. Briefly, each participant carried a personal monitor to collect two 12-h air samples and gave a breath sample at the end of the sampling period. Breath samples were collected using a specially designed spirometer and cartridges containing Tenax GC adsorbent to trap the organic vapors for later analysis. These cartridges, which constituted the subset of samples for the present study, were analyzed in our laboratory using a thermal desorption technique followed by combined GC–MS.

Data processing procedures

Using a Digital PDP-11/34 minicomputer and the RSX11M operating system, the raw GC–MS data files were evaluated with an automated spectrum-enhancement algorithm (program CLEANUP) that automatically locates components to yield a set of pure spectra that are free of background contributions and peaks from overlapping

components[24]. Peak matching and file averaging for each sample group were carried out using two more programs (TIMSEK and MAKLOC)[25]. The final reduced data set, consisting of average retention indices and relative concentrations for the components in each sample group, was transmitted to an IBM-compatible personal computer (PC) via an RS-232C serial ASCII interface.

The 26 smokers and 43 non-smokers gave a total of 69 GC–MS profiles and 230 compounds, which served as the primary data base. Because the frequency distributions of exhaled breath values for the TEAM compounds are closer to log normal than normal[15], we applied the Fisher exact test and Mann–Whitney U statistics from the BMDPC personal-computer software package[26] to the individual compounds to yield a subset of GC–MS peaks for further analysis. Principal component analysis was used to further reduce the dimensionality of the data, and discriminant analysis provided a suitable classification model. The peaks were identified by comparing their "clean" mass spectra with standard spectra in the PC-based CD-ROM (Compact Disk-Read Only Memory) edition of the Registry of Mass Spectral Data, using the Probability-Based-Matching search program[27].

RESULTS AND DISCUSSION

Broad-spectrum GC–MS analysis of the 69 breath samples identified close to 230 compounds. The average number of compounds identified in the 26 smokers samples was 144 ($\pm$27) and 129 ($\pm$28) in the 43 non-smokers samples.

Data preparation

The CLEANUP–TIMSEK–MAKLOC algorithms[24,25] provide a powerful and convenient means of enhancing and compositing large groups of complex raw GC–MS data files. The programs produce summary lists of the average relative concentrations and effective retention indices for each sample group of interest. These lists can be used directly to analyze the data further.

The CLEANUP program was used to improve the quality of the mass spectral data by automatically locating and extracting components to produce a set of pure spectra that are free of background contributions and overlapping peaks. Once a clean spectrum was defined by the program, its area was calculated and used to evaluate the relative concentration of the component in the sample by comparing the peak area of that component with the peak area of a standard. The external standard, perfluorobenzene, was added to each Tenax GC cartridge just before analysis.

To match peaks and generate composite GC–MS profiles for non-smokers and smokers, each peak in the profiles was first subjected to retention time scaling (program TIMSEK) to compensate for variations in retention time between individual experiments. The retention times were scaled by converting the raw chromatographic retention data (spectrum scan numbers) into a set of standardized retention indices relative to a set of equally spaced marker peaks that occurred throughout the data base and were easily located. These peaks were initially used to create a reference calibration file that related spectrum scan number to peak position. For all subsequent profiles, the reference peaks were used to calculate an effective retention index scale for the entire profile. The retention scaling program also calculated relative concentrations (*i.e.*, peak area ratios) for each component with respect to the standard.

The peak matching program (program MAKLOC) takes "clean" sets of mass spectral and effective retention index data for each component in a GC–MS profile and makes peak-by-peak comparisons with the corresponding data in associated profiles. To do this, one of the profiles of interest was set up as an historical (standard) library. The program then defined successive time windows, in each of which the unknown spectrum was compared with every reference spectrum in the window of the historical library by performing a relative retention index (RRI) match and a spectrum match. Matching effectiveness was expressed in terms of a score that gave a measure of spectral similarity and RRI proximity. When a satisfactory match occurred, the peak was added to the historical library, and the relative concentration and RRI data for the entry were updated.

The composited data files, which include statistics on the frequency of occurrence and the reproducibility of the average relative concentrations in the two sample groups of interest, were transferred from the PDP-11/34 minicomputer to a spreadsheet in the PC. For those compounds not detected, values were taken to be the average estimated limit of detection. A complete summary of the group data for the smokers' and non-smokers' samples is available on request from the author. The data show the presence of many different chemical classes in the sample groups, including aliphatics, aromatics, halogenated hydrocarbons, alcohols, aldehydes, ethers, esters and ketones. A number of compounds occurred with a frequency of 100% in one of the groups and at least 90% frequency in the other. Of especial interest was the fact that three compounds (1-penten-3-yne, a methyl-1,3-cyclopentadiene isomer and 2,5-dimethyl furan) occurred with high frequency in the breath of the smokers but were entirely absent from, or only occasionally observed in, the non-smokers' breath. Thus, they could serve as marker compounds to distinguish smokers from non-smokers.

Statistical data analysis: peak selection

Given the large number of constituents of human expired air and their complex interrelationships[13], multivariate data analysis provides a more realistic and objective means of extracting useful information from the GC–MS profiles than do piecemeal univariate procedures. However, certain problems complicate the determination of significant differences in the constituents of human expired air between groups of smokers and non-smokers. First, the number of breath samples available for the study is small relative to the number of constituents detected. Second, although multivariate statistical techniques can be effective, they can also identify fortuitous associations that are neither reproducible nor physically meaningful[28]. Generally, it has been shown that, as long as the ratio of the number of samples in the data set (*i.e.*, the GC–MS profiles) to the number of descriptors (*i.e.*, the individual peaks) per sample is greater than 3, then the probability of achieving complete separation due to chance alone is low. Therefore, no more than 23 peaks may be analyzed by multivariate statistics to identify characteristic compounds in the breath of smokers.

For each of the 230 compounds in the smokers and non-smokers samples, χ^2 statistics (Fisher exact test) were used in the univariate case to test smoker–non-smoker differences in terms of the presence or absence of a particular component. A type I error rate of $P < 0.05$ was taken to indicate significance in this initial screen. Then, Mann–Whitney U statistics were computed to test for significant differences in the relative concentrations of each compound between the two groups. This procedure

TABLE II

POTENTIALLY SIGNIFICANT PEAKS BASED ON FISHER EXACT AND MANN–WHITNEY TESTS

Peak No.	*RRI*	*Formula*	*Compound*	*Fisher Exact test*			*Mann–Whitney test*
				Occurrence (%)		*P value*	
				Non-smokers	*Smokers*		*P value*
1	472	C_3H_6O	Acetone	98	88	0.1466	0.0110
2	531	C_5H_6	1-Penten-3-yne	5	77	0.0000	0.0000
3	550	C_5H_8	Cyclopentene	21	62	0.0008	0.0006
4	572	$C_4H_6O_2$	2,3-Butane dione	35	54	0.0979	0.0313
5	577	C_4H_8O	2-Methyl propanal	53	31	0.0553	0.0398
6	590	C_6H_{12}	1-Hexene	74	92	0.0596	0.0000
7	610	C_6H_{12}	Methyl pentene isomer	21	88	0.0000	0.0000
8	622	C_6H_{12}	Methyl pentene isomer	40	88	0.0000	0.0001
9	636	C_6H_8	Methyl-1,3-cyclopentadiene isomer	0	77	0.0000	0.0000
10	640	C_6H_8	Methyl-1,3-cyclopentadiene isomer	0	62	0.0000	0.0000
11	648	C_6H_{10}	Methyl cyclopentene isomer	9	54	0.0001	0.0000
12	649	C_6H_{10}	3-Methyl cyclopentene	19	54	0.0029	0.0006
13	652	C_6H_6	Benzene	100	100	–	0.0716
14	688	C_2HCl_3	Trichloroethylene	81	73	0.3016	0.0393
15	697	C_6H_8O	2,5-Dimethyl furan	0	92	0.0000	0.0000
16	732	C_7H_{16}	Alkane (tentative identification)	51	81	0.0125	0.0345
17	736	C_7H_{10}	Methyl-1,3,5-hexatriene isomer	0	58	0.0000	0.0000
18	789	$C_5H_{12}O$	3-Methyl-1-butanol	16	77	0.0000	0.0000
19	853	C_8H_{10}	Ethyl benzene	100	100	–	0.3187
20	861	C_8H_{10}	*m/p*-Xylene	100	100	–	0.2546
21	878	C_8H_8	Styrene	77	96	0.0305	0.0146
22	884	C_8H_{10}	*o*-Xylene	100	100	–	0.1422

tests the null hypothesis of no difference in the mean concentrations of a compound between the two sample sets. Again, a type I error rate of $P < 0.05$ was assumed to be significant.

Those peaks which were significant by one or more of the tests used, and occurred in at least one of the two sample groups with a frequency greater than 50%, were selected to serve as a starting point for further analysis. The nineteen peaks that satisfied these requirements are listed in Table II. They include the aromatics benzene and styrene, which were found in earlier TEAM studies to be significantly higher in the breath of smokers compared to non-smokers[16,17]. Table II also includes the compounds ethyl benzene and the xylene isomers which, while not satisfying our peak selection requirements, nevertheless exhibited the same behaviour as benzene and styrene in the earlier work.

Multivariate statistical analysis

Before attempting to reduce the dimensionality of the data, a logarithmic transformation was applied to each relative concentration, which substantially reduced the relatively large variance in the raw peak area values. Factor analysis and discriminant analysis were performed on the 22 selected peaks. Factor analysis was

used to rank the peaks according to their importance in describing the data variance, while discriminant analysis served to obtain the optimal separation between the two sample groups from the minimum number of peaks[29,30].

Factor analysis attempts to identify underlying patterns in the samples by describing the variance in the data in as few factors as possible, while still retaining the information content of the originals. The observed variables are placed into linear combinations, and the first factor is the combination that accounts for the largest amount of variance in the sample. Successive factors explain progressively smaller fractions of the total sample variance.

In order to determine the minimum number of factors needed to represent the data, we computed the total variance explained by each factor as well as the cumulative percentage. The results are summarized in Table III. Almost 70% of the total variance is attributable to the first four factors, the remaining eighteen factors together accounting for only 30% of the variance. It therefore appears that the data may be adequately represented by a model based only on the first four factors.

Although the first few factors explain the major portion of the variance, they do not identify the most important individual peaks. To facilitate this identification, we performed factor analysis rotation using the varimax method, which minimizes the number of variables with high loadings on a factor, thus transforming the initial matrix into one that is easier to interpret[29]. Table IV shows the rotated factor loadings for the first four factors. The factor loading matrix has been sorted so that variables with high

TABLE III

TOTAL VARIANCE EXPLAINED BY EACH FACTOR

Factor	*Variance explained*	*Percentage of total variance*	*Cumulative percentage*
1	8.528	38.8	38.8
2	3.902	17.7	56.5
3	1.576	7.2	63.7
4	1.382	6.2	69.9
5	1.086	5.0	74.9
6	1.018	4.6	79.5
7	0.869	4.0	83.5
8	0.670	3.0	86.5
9	0.483	2.2	88.7
10	0.408	1.9	90.6
11	0.404	1.8	92.4
12	0.321	1.5	93.9
13	0.263	1.2	95.1
14	0.237	1.0	96.1
15	0.198	0.9	97.0
16	0.172	0.8	97.8
17	0.140	0.6	98.4
18	0.113	0.6	99.0
19	0.088	0.4	99.4
20	0.061	0.2	99.6
21	0.041	0.2	99.8
22	0.039	0.2	100.0

TABLE IV

ROTATED FACTOR LOADING MATRIX (VARIMAX METHOD)

Peak No.	*RRI*	*Factor*			
		1	*2*	*3*	*4*
17	736	0.928			
10	640	0.874			
2	531	0.861		0.323	
9	636	0.854			
15	697	0.751			
11	648	0.675			−0.381
7	610	0.660		0.486	
6	590	0.600			
22	884		0.916		
19	853		0.915		
20	861		0.873		
13	652		0.736		
3	550	0.358		0.857	
12	649	0.273			0.911
16	732				
5	577				
18	789	0.416			
1	472				
4	572	0.326			
14	688		0.341		
8	622	0.310	0.303	0.299	
21	878		0.519		

loadings on the same factor appear together. Small factor loadings (<0.250) are omitted from the table. Several peaks have large loadings on the same factor, *e.g.*, 17, 10, 9, 15 and 6 on the first factor; 22, 19, 20 and 13 on the second factor, etc. In addition, certain peaks (*e.g.*, 7, 3, 12 and 8) share information over more than one factor. The first factor contains many of the peaks that occurred often in smokers' breath and rarely in non-smokers' breath. The second factor includes all of the aromatic compounds previously found to be at higher concentrations in smokers' breath[16,17].

The minimum number of peaks needed to distinguish between smokers' and non-smokers' samples was ascertained by discriminant analysis, using the peaks determined from the factor technique. In discriminant analysis, group membership of the samples is defined from the beginning, and the variables are linearly combined in such a way that the defined groups are as statistically distinct as possible[26,29,30]. Variables are entered into, or removed from, the analysis on a stepwise basis and the discriminant function maximizes the ratio of the inter-group to intra-group variance at each step. We constructed discriminant functions using the information in Table IV to maximally differentiate smokers from non-smokers, then tested the accuracy of each estimated function both statistically and by examining the concordance between the observed and estimated classifications. A classification accuracy of greater than 80% was considered to be significant.

TABLE V

CLASSIFICATION RESULTS FROM STEPWISE DISCRIMINANT ANALYSIS USING PEAKS HEAVILY LOADED ON A SINGLE FACTOR

Group	*% Correct*[a]	*Number of samples*	*Number misclassified*
Non-smokers	100	43	0
Smokers	89	26	3
Overall:	96		

[a] Peaks 10 and 15 included in discriminant function.

Ideally, the validation of each result should be based on the classification of an independent data set. However, no extra samples were available for independent evaluation and the data set used was not large enough to be randomly split into two groups, one to derive the discriminant function and the other to test it. Consequently, we used the "jackknife" validation procedure in order to estimate misclassification rates and minimize bias[26,30]. In this "leave-one-out" method, each of the samples is left out in turn, the discriminant function is recalculated based on the remaining $(n-1)$ samples, and then the left-out sample is classified. This procedure continues until all samples in the set have been left out and classified once.

Following the approach described by Parrish *et al.*[31], we used the information in Table IV to compare the effect on the overall classification of peaks that are heavily loaded on only one factor as opposed to peaks that share information. Although several peaks were designated from the first category, namely 9, 10, 15 and 17 (factor 1), and 19, 20 and 22 (factor 2), only two (peaks 10 and 15) were selected in the stepwise discriminant analysis. Examination of the data in Table II reveals that peak 15 occurs with high frequency in the smokers' samples while both are completely absent from the non-smokers' samples. Peak 15 exhibits a disproportionate power to discriminate in this case, as indicated by the calculated F-statistic to enter or remove a variable from the discriminant function ($F = 290$ for peak 15; $F = 3.5$ for peak 10)[29,30]. The classification results from the analysis are shown in Table V. The overall percent of samples correctly classified was 96% for both the discriminant model and the jackknifed classification. All of the misclassification involves the smokers group, in which 23 samples are predicted correctly to be members of this group, while three are assigned to the non-smokers group. Further evidence of the dominant role played by peak 15 in discriminating between the two groups is provided by the fact that the same classification results are obtained when only this peak is entered into the analysis.

To establish the discriminatory importance of the remaining peaks heavily loaded on one factor, peak 15 was removed from contention and the discriminant function was recalculated using peaks 9, 10, 17, 19, 20 and 22. The results are summarized in Table VI and show that a jackknifed classification accuracy of 90% was obtained with peaks 9, 10, 20 and 22 selected in the analysis. In this case, the calculated F-values indicate that peak 9 plays a dominant discriminating role followed, to a much lesser extent, by peak 20.

Stepwise discriminant analysis was also performed using peaks that share information, in addition to the peaks that have a high loading on only one factor. The

TABLE VI

SUMMARY OF PEAKS EVALUATED AND JACKKNIFED CLASSIFICATION RESULTS FROM STEPWISE DISCRIMINANT FUNCTION ANALYSIS

Analysis criterion	*Case*	*Peaks considered*[a]	*Peaks selected*[b] *(F-value in parentheses)*	*Overall percentage correctly classified*
High loaded single-factor peaks	1	9, 10, 15, 17, 19, 20, 22	15(290), 10(3.5)	96
	2	9, 10, 17, 19, 20, 22	9(115), 20(5.9), 10(1.3), 22(1.1)	90
High loaded single-factor plus cross-term peaks	3	2, 3, 6, 7, 9, 10, 11, 12, 13, 15, 17, 19, 20, 21, 22	15(290), 10(3.5), 2(1.7), 7(1.0)	96
	4	2, 3, 6, 7, 9, 10, 11, 12, 13, 17, 19, 20, 21, 22	9(115), 7(5.4), 20(5.2), 2(4.8), 17(3.9), 10(2.1), 3(1.7), 22(1.2)	90
Major discriminating peaks selected from cases 1–4	5	2, 3, 7, 9, 10, 15, 17, 20, 22	15(290), 10(3.5), 2(1.7), 7(1.0)	96
Aromatics[c]	6	13, 19, 20, 21, 22	21(7.4), 13(4.1)	61

[a] From data in Table IV (peak numbers identified in Table II).

[b] Using values of 1.0 for *F*-to-enter and remove peaks in BMDPC 7M discriminant analysis program.

[c] Identified in earlier TEAM studies[16,17].

classification results obtained are summarized in Table VI (case 3). Although 15 peaks were considered in the analysis and 4 were selected, peak 15 again displayed an exaggerated power to discriminate and gave an overall classification accuracy of 96%. As before, peak 15 was removed from consideration in order to establish the relative importance of the other compounds, and the discriminant function was recalculated using the remaining 14 peaks. The results are shown in Table VI (case 4) and again reveal the discriminatory importance of peak 9, followed this time by peak 7.

The major discriminating peaks resulting from the analyses described above were thereupon chosen for evaluation, and the results are summarized in Table VI (case 5). Once again, the dominating influence of peak 15 is clearly evident, giving the same classification accuracy (96%) obtained in all of the earlier analyses in which it was one of the peaks considered.

Finally, the discriminant technique was applied to the two groups, using the aromatic peaks (benzene, styrene, ethyl benzene and the xylene isomers) which were found in earlier TEAM studies to be significantly higher in the breath of smokers[16,17]. When the peaks were entered into the analysis, only styrene was selected, and the overall percentage correctly classified decreased to 61%. This suggests that these compounds contain less discriminating information than the compounds discussed above.

Significance of selected compounds

Overall, the breath components which gave the best discrimination between the smokers and non-smokers samples, under the various conditions that were considered, were 2,5-dimethyl furan (peak 15), the methyl-1,3-cyclopentadiene isomers (peaks 9 and 10), *m*/*p*-xylenes (peak 20) and a methyl pentene isomer (peak 7). Of these, 2,5-dimethyl furan plays a dominant role in distinguishing between the two groups.

This compound (2,5-dimethyl furan) has been identified as a major gas-phase constituent of tobacco smoke[32]. It probably arises because of the rich oxygen environment in which the combustion of tobacco takes place, which results in the occurrence of a large number of oxygenated compounds in mainstream smoke. The compound was also previously identified in exhaled breath samples with a 60% occurrence frequency, but was not observed in the corresponding breathing-zone air samples[33]. However, its potential usefulness as a marker of cigarette exposure could not be evaluated from this work, since no distinction was made between smokers and non-smokers.

CONCLUSIONS

The compound 2,5-dimethyl furan and the other major discriminating compounds identified in the present study illustrate the striking power of factor analysis and discriminant analysis to differentiate with a high degree of accuracy between smokers and non-smokers, on the basis of the volatile compounds present in their exhaled breath. Thus, they appear to be effective biochemical markers of smoking, and suggest that the analysis of exhaled breath could provide a reliable non-invasive method for mass screening studies. However, the half-lives of these compounds must be determined before they can be exploited as practical indicators of exposure to smoking. A simple two-parameter time-dependent model has been shown to have some success in estimating biological half-lives of several VOCs[34], and recently a "washout" study was performed over a 10-h period in a pure air chamber which resulted in measured half-lives for tetrachloroethylene and chloroform[35]. Similar measurements are needed on the compounds identified here as potential markers of exposure to cigarette smoke so that a suitable model relating such exposure to body burden can be developed.

ACKNOWLEDGEMENTS

The author thanks Dr. Lance A. Wallace of the U.S. Environmental Protection Agency for providing information from the activity screeners on the smoking habits of the subjects in the study, and Drs. Demetrios J. Moschandreas, Robert G. Gibbons and Lance A. Wallace for reviewing the manuscript. This investigation was supported by PHS grant number 1 R03-CA43959-01, awarded by the U.S. Department of Health and Human Services.

REFERENCES

1 *Surgeon General Report: The Health Consequences of Smoking—Cancer*, U.S. Department of Health and Human Services, Washington, DC, 1982.

2 P. Greenwald and J. W. Cullen, *J. Natl. Can. Inst.*, 74 (1985) 543.

3 M. F. Dube and C. R. Green, *Recent Adv. Tob. Sci.*, 8 (1982) 42.
4 T. M. Vogt, S. Selvin, G. M. Widdowson and S. B. Hurley, *Am. J. Public Health*, 67 (1977) 545.
5 N. Hengen and M. Hengen, *Clin. Chem.*, 24 (1978) 50.
6 P. Hill, N. J. Haley and E. L. Wynder, *J. Chron. Dis.*, 36 (1983) 439.
7 N. J. Haley, C. M. Axelrod and K. A. Tilton, *Am. J. Public Health*, 73 (1983) 1204.
8 N. J. Haley and D. Hoffmann, *Clin. Chem.*, 31 (1985) 1598.
9 R. Pojer, J. B. Whitfield, V. Poulos, I. F. Eckhard, R. Richmond and W. J. Hensley, *Clin. Chem.*, 30 (1984) 1377.
10 M. J. Jarvis, in I. K. O'Neill, K. D. Brunnemann, B. Dodet and D. Hoffmann (Editors), *Environmental Carcinogens—Methods of Analysis and Exposure Measurement, Vol. 9, Passive Smoking, (IARC Scientific Publications*, No. 81), International Agency for Research on Cancer, Lyon, 1987, pp. 43–58.
11 H. Muranaka, E. Higashi, S. Itani and Y. Shimizu, *Int. Arch. Occup. Environ. Health*, 60 (1988) 37.
12 B. K. Krotoszynski, G. Bruneau and H. J. O'Neill, *J. Anal. Toxicol.*, 3 (1979) 225.
13 A. Manolis, *Clin. Chem.*, 29 (1983) 5.
14 M. Berlin, J. C. Gage, B. Gullberg, S. Holm, P. Knutsson and A. Tunek, *Scand. J. Work Environ. Health*, 6 (1980) 104.
15 L. A. Wallace, E. D. Pellizzari, T. D. Hartwell, C. M. Sparacino, L. S. Sheldon and H. Zelon, *Atmos. Environ.*, 19 (1985) 1651.
16 L. A. Wallace and E. D. Pellizzari, *Toxicol. Lett.*, 35 (1986) 113.
17 L. A. Wallace, E. D. Pellizzari, T. D. Hartwell, R. Perritt and R. Ziegenfus, *Arch. Environ. Health*, 42 (1987) 272.
18 S. M. Gordon, J. P. Szidon, B. K. Krotoszynski, R. D. Gibbons and H. J. O'Neill, *Clin. Chem.*, 31 (1985) 1278.
19 L. A. Wallace, *Report EPA 600/6-87/002a, Total Exposure Assessment Methodology (TEAM) Study: Summary and Analysis*, Vol. I, U.S. Environmental Protection Agency, Washington, DC, 1987.
20 E. D. Pellizzari, K. Perritt, T. D. Hartwell, L. C. Michael, R. Whitmore, R. W. Handy, D. Smith and H. Zelon, *Report EPA 600/6-87/002b, Total Exposure Assessment Methodology (TEAM) Study: Elizabeth and Bayonne, New Jersey; Devils Lake, North Dakota; and Greensboro, North Carolina*, Vol. II, U.S. Environmental Protection Agency, Washington, DC, 1987.
21 E. D. Pellizzari, K. Perritt, T. D. Hartwell, L. C. Michael, R. Whitmore, R. W. Handy, D. Smith and H. Zelon, *Report EPA 600/6-87/002c, Total Exposure Assessment Methodology (TEAM) Study: Selected Communities in Northern and Southern California*, Vol. III, U.S. Environmental Protection Agency, Washington, DC, 1987.
22 R. W. Handy, D. J. Smith, N. P. Castillo, C. M. Sparacino, K. Thomas, D. Whitaker, J. Keever, P. A. Blau, L. S. Sheldon, K. A. Brady, R. L. Porch, J. T. Bursey and E. D. Pellizzari, *Report EPA 600/6-87/002d, Total Exposure Assessment Methodology (TEAM) Study: Standard Operating Procedures*, Vol. IV, U.S. Environmental Protection Agency, Washington, DC, 1987.
23 R. W. Whitmore, *Atmos. Environ.*, 22 (1988) 2077.
24 R. G. Dromey, M. J. Stefik, T. C. Rindfleisch and A. M. Duffield, *Anal. Chem.*, 48 (1976) 1368.
25 D. H. Smith, M. Achenbach, W. J. Yeager, P. J. Anderson, W. L. Fitch and T. C. Rindfleisch, *Anal. Chem.*, 49 (1977) 1623.
26 *BMDPC: Guide to Using BMDP on the IBM PC*, BMDP Statistical Software, Los Angeles, CA, 1987.
27 *1987 Registry of Mass Spectral Data, CD-ROM Edition*, Wiley Electronic Publishing, New York, 1987.
28 B. K. Lavine, P. C. Jurs, D. R. Henry, R. K. Vander Meer, J. A. Pino and J. E. McMurry, *Chemom. Intell. Lab. Syst.*, 3 (1988) 79.
29 M. J. Norusis, *SPSS/PC+Advanced Statistics for the IBM PC/XT/AT*, SPSS, Chicago, IL, 1986.
30 D. D. Wolff and M. L. Parsons, *Pattern Recognition Approach to Data Interpretation*, Plenum Press, New York, 1983.
31 M. E. Parrish, B. W. Good, M. A. Jeltema and F. S. Hsu, *Anal. Chim. Acta*, 150 (1983) 163.
32 C. E. Higgins, W. H. Griest and G. Olerich, *J. Assoc. Off. Anal. Chem.*, 66 (1983) 1074.
33 L. A. Wallace, E. Pellizzari, T. Hartwell, M. Rosenzweig, M. Erickson, C. Sparacino and H. Zelon, *Environ. Res.*, 35 (1984) 293.
34 L. A. Wallace, E. Pellizzari, T. Hartwell, H. Zelon, C. Sparacino, R. Perritt and R. Whitmore, *J. Occup. Med.*, 28 (1986) 603.
35 S. M. Gordon, L. A. Wallace, E. D. Pellizzari and H. J. O'Neill, *Atmos. Environ.*, 22 (1988) 2165.