2,5-dimethylfuran as a validated biomarker of smoking status

Mar Castellanos, MD, PhD [1,2]; Rosa Suñer, PhD [3]; José M. Fernández-Real, MD, PhD [4]; Juan M. Sanchez Navarro, PhD [2,5,*]

1. Department of Neurology, Complejo Hospitalario Universitario de A Coruña (CHUAC), A Coruña, Spain. Maria.del.Mar.Castellanos.Rodrigo@sergas.es

2. Instituto de Investigación Biomédica de A Coruña (INIBIC), A Coruña, Spain

3. Department of Nursing, University of Girona, Girona (Spain). rosa.sunyer@udg.edu

4. Department of Diabetes, Endocrinology and Nutrition, Dr Josep Trueta University Hospital, Girona, Spain. jmfreal@idibgi.org

5. Department of Chemistry, University of Girona, Girona, Spain. juanma.sanchez@udg.edu

* Correspondence author

Correspondence to: Juan M. Sanchez
Department of Chemistry
Science Faculty
University of Girona
Aurèlia Capmany, 69
Girona – 17003
Spain
Phone: (+34) 636569984
E-mail: juanma.sanchez@udg.edu

Abstract

Introduction: Exposure biomarkers are required in tobacco use studies to accurately assess smoking status since self-reporting usually results in misclassification estimates. This study uses breath analysis and assesses some volatile organic compounds (VOCs) as potential biomarkers of tobacco smoke exposure.

Methods: Forced-expiratory breath samples were obtained from 377 volunteers (174 smokers and 203 non-smokers). Exhaled breath levels of different VOCs previously related to tobacco smoke were evaluated. The toluene-to-benzene ratio was evaluated as this ratio has been found to be different in atmospheric samples and tobacco smoke emissions. Finally, breath analyses from 64 patients attending a clinical practice were evaluated and the results were compared to their self-reporting status.

Results: Univariate analysis shows that all compounds evaluated gave significant differences (*p*<0.001). Receiver operating characteristic (ROC) curves suggest that xylenes and toluene are not able to accurately determine smoking status, and benzene and the T/B ratio present potential utility in certain conditions. The highest discriminant capacity was obtained for 2,5-dimethylfuran (AUC=0.982, 95% CI: 0.969-0.995), with a cut-off value of 0.016 ppbv (sensibility=0.965, specificity=0.896). Drinking coffee was the only confounding parameter that can give low breath levels for this compound. The evaluation of the results obtained from the patients attending a clinical practice showed that 8% of people who claim to be non-smokers hid their real smoking status.

Conclusions: The results obtained confirm that the determination of 2,5-dimethylfuran in breath samples is a good and simpler alternative to conventional blood or urine tests for assessing smoking status.

Implications: Analysis of 2,5-dimethylfuran in breath samples results in a simple and fast method for the determination of the smoking status of a person. This methodology presents multiple advantages as it is neither invasive nor embarrassing for patients attending clinical practices. Moreover, analysis of biomarkers in breath samples is simpler and faster than using conventional methods based on urine or blood analysis.

INTRODUCTION

Smoking is associated to many adverse health effects [1,2] and also has a clear economic effect on the cost of health services [3,4]. For this reason, many life insurance companies offer significant premium reductions to non-smokers. In the U.S., the Affordable Care Act (ACA) allows insurance companies to charge smokers up to 50% more than non-smokers through a tobacco surcharge [5-7].

Accurate assessment of smoking status is critical for determining tobacco use, population risk, and smoking diseases. Self-reported smoking is widely used to determine exposure to tobacco smoke and estimate prevalence of cigarette smoking. However, this procedure tends to underestimate the true prevalence of smoking when a biological sample has also been analyzed for comparison [8]. Although self-report is a fair estimator of smoking prevalence, misclassification rates for non-smokers ranging from 1% to 10% have typically been reported [8-14]. These misclassification rates tend to be still larger in ex-smokers, pregnant women, and clinic-based studies [9,15], which suggests that people may deny their current smoking habits due to social stigma or fear, particularly in situations where they may have been given advice to stop smoking by a doctor.

The WHO Study Group on Tobacco Product Regulation [16] concluded that exposure biomarkers are required to support exposure reduction claims in studies defining the dependence potential of different products and in evaluating the effects of specific regulatory changes on exposures in the general population. Therefore, it is necessary to find a biomarker of tobacco exposure to confirm self-reported information.

Nicotine is the main addictive component in tobacco products, a major constituent of cigarettes, and is highly specific to tobacco smoke. However, its short half-life time in biological fluids ($t_{1/2}$=2-3 h in blood) makes it necessary to determine a nicotine metabolite. In the case of tobacco smoke, serum cotinine ($t_{1/2}$=15-19 h in blood, urine and saliva), the major metabolite of nicotine, is considered to be the standard for measuring exposure [15,17]. However, blood analysis is an invasive method to obtain the samples. Urine cotinine or its glucuronide conjugate are reliable measures of nicotine uptake and are commonly used biomarkers of tobacco exposure [15,17], although it should be noted that total cotinine may also reflect nicotine exposure from tobacco substitutes, such as nicotine patches [18] and nicotine chewing gum [19]. Tobacco-specific nitrosamines, particularly 4-(methylnitrosamino)-1-3-(pyridyl)-1-butanone (NNK) and its major metabolite 4-(methylnitrosamino)-1-3-(pyridyl)-1-butanol

(NNAL), are more accurate urinary biomarkers of tobacco exposure [20]. Although urine measurement is not invasive, it can be viewed as psychologically invasive or embarrassing, there is a biological hazard involved in specimen handling, and it may be hard to apply in a pediatric setting or in the case of large numbers of study subjects. Other non-invasive methods such as saliva, hair, and sputum analysis have been used to evaluate tobacco exposure [15,17].

The least invasive and probably the simplest method to perform this study is through breath analysis. Exhaled CO has normally been used to assess recent tobacco exposure (<8h) [17,21] despite there being considerable inter-individual variability. Tobacco smoke is an aerosol produced by complex and overlapping burning-, pyrolysis-, pyrosynthesis-, distillation-, sublimation- and condensation-processes during the smoking of cigarettes, which comprises a highly complex chemical mixture of non-specific products of organic material combustion and chemicals that are specific to the combustion of tobacco and other components of the cigarette [22]. Approximately 5% (w/w) of mainstream smoke is composed by volatile organic compounds (VOCs), which are formed by the incomplete combustion of tobacco during and between puffs. Previous studies have demonstrated that active smoking increases the levels of different VOCs in breath and blood [23-30], and active smokers can be discriminated by higher values for combustion products such as furans, as well as benzene, toluene and xylene aromatic hydrocarbons. It has been reported that 2,5-dimethylfuran plays a dominant role in distinguishing between smokers and non-smokers [25,27,29,30].

A preliminary study [29] found that 2,5-dimethylfuran is a highly selective breath biomarker of smoking status as this compound is able to differentiate between social smokers and non-smokers. In the present study, we have performed breath analysis of a large cohort of smokers and non-smokers to determine the validity of this technique as an alternative to conventional blood and urine analysis for the determination of smoking status. The cut-off value for the compound with the highest discriminant capacity for smoking status has been determined and the proposed method has been used to check the validity of self-reports at two different medical practices.

MATERIALS AND METHODS

Subjects

387 adult volunteers participated in the study. Before taking breath samples, participants were informed of the nature of the test and the aims of the study. Inclusion

criteria were that a person was considered to be a smoker when he/she admitted to a smoking habit of at least one cigarette/day and had smoked within the previous 24 h. Within this group, only cigarette smokers were evaluated and the exclusive use of any other tobacco product, such as e-cigarettes, was an exclusion criterion. Given that a previous study has indicated that some VOCs can detect smoking status after more than 24 h without smoking [29], we decided to exclude ten participants from the statistical calculations as they reported being social smokers, consuming less than one cigarette/day on average, and that more than 24 h had gone by since they last smoked a cigarette. Of the 377 adult volunteers included in the study, 127 were men (33.7 %) and 250 were women (66.3 %), and the mean age was 29.2 years (range 16-61). No requirements related to food and drink ingestion were made prior to breath sampling, although the volunteers were asked whether they had drunk coffee in the previous hours because it has been reported that 2,5-dimethylfuran can be released by roasted coffee beans [31,32]. Thirty-six volunteers admitted to the use of cannabis and 29 of these mixed cannabis with tobacco.

Under the conditions indicated for being considered smokers, 174 volunteers were included in this group. The smoking habits of these subjects were recorded: 20 people (11.5%) reported smoking >20 cigarettes/day, 48 (27.6%) smoked between 10 and 20 cigarettes, 31 (17.8%) around 10 cigarettes, and 75 (43.1%) less than 10. With regards to the time since the last cigarette smoked: breath samples were collected just after smoking (<5 min after the last cigarette) in the case of 49 smokers (28.2%), 30 of them (17.2%) reported a time span of between 5 and 30 min, 28 smokers (16.1 %) gave their samples between 30 and 60 min after smoking, 43 (24.7%) gave samples between 1 and 6 hours after smoking, and 24 (13.8 %) reported that they had not smoked since the previous day (10-15 h time span).

For the validation of the proposed results, breath samples from 64 people (43 females; mean age 42.5 years, range 23-61) attending a neurological and a endocrinology practice were assessed. Smoking status was taken during these visits and the information self-reported by patients was compared with the results obtained from breath analyses. When disagreement was observed, patients were contacted again to confirm that their initial self-report was not correct.

Breath analysis

Different VOCs were selected for their evaluation as smoking biomarkers, taking into account preliminary results [23-30]. Benzene was evaluated given that it is the VOC that has most frequently been proposed as a smoking biomarker in the literature. Toluene and xylenes were chosen because VOC emissions by cigarette smoke are usually dominated by benzene and these compounds [22,33-36]. 2,5-dimethylfuran was included as recent studies have demonstrated its strong correlation with smoking status [25,27,29].

For the analysis of breath samples, an "in-house" capillary thermal desorption device connected to a gas chromatograph with mass spectrometry detection (GC-MS) (Thermo Scientific, Waltham, MA, USA) was used [37,38]. The microtrap used in this study has been specifically developed for the analysis of VOCs in breath samples at ppbv-pptv levels. Specific details about trap design, the GC-MS method and its validation are given in the Supplementary Materials.

Forced-expiratory breath samples were collected for each individual as follows: the first 2–3 s of the expiration were not collected in order to minimize the sampling of dead-space air, and the remaining fraction was collected until about 900 mL of breath had been introduced into a cleaned 1 L Tedlar gas-sampling bags (SKC Inc., Eighty Four, PA, USA). Each sample was analyzed no more than two hours after being collected to avoid the loss of analytes from the bags [39]. For each sample, 750 cm^3 of breath were required for the chromatographic analysis (i.e. breath samples were moved through the microtrap during 25 min at a fixed flow rate of 30 $cm^3 \cdot min^{-1}$).

Each Tedlar bag was cleaned with purified nitrogen several times before new samples were collected. In order to confirm the validity of the cleaning process, the last portion of nitrogen collected in the cleaning process was analyzed in the same conditions as breath samples to confirm that no detectable levels of any target compound were found.

Statistical Analysis

Statistical analysis was performed using SPSS for Windows Version 15.0. For calculations of statistical significance, two-sided testing was used and p<0.05 was considered as significant. Shapiro-Wilk and Kolmogorov-Smirnov tests were used to study the distribution of the compounds evaluated in the samples. The results indicated that the chosen analytes do not follow a normal distribution neither in the case of smokers nor non-smokers (p<0.001). Continuous variables are expressed as median [quartiles] and the Mann-Whitney U-test was used to compare the values found between smokers and non-smokers. Receiver operating characteristic (ROC) curves were used to assess the discriminant power of each individual compound and to determine the best cut-off value for the prediction of smoking status. Multivariate logistic regression analysis was used to determine the compounds that can predict smoking status.

RESULTS

Two-hundred and three (53.8%) of the 377 subjects included in the study reported being non-smokers, whereas 174 were smokers (46.2%). No differences in smoking status were found by sex as 72 of the 127 men were non-smokers (56.7%) and 55 smokers (43.3%), whereas 131 of the 250 women were non-smokers (52.4%) and 119 were smokers (47.6%). Toluene was the only VOC that gave a small but significant difference between sexes (*p*=0.038), with higher values detected in women (3.215 [1.215-6.872] ppbv) than in men (1.986 [1.038-5.738] ppbv).

Benzene, toluene and xylenes were detected in all samples evaluated, both in smokers and non-smokers; only *o*-xylene was not detected in the sample of one non-smoker (0.5%), who had not smoked in the previous 15 h. 2,5-dimethylfuran was detected in 172 smoker samples (98.9%) and was not detected in 173 non-smokers (85.2%). Table 1 shows the results obtained in the determination of the target analytes in smokers and non-smokers.

In the univariate analysis (Mann-Whitney U-test), significantly higher values were detected for all compounds in the smoker group, which agrees with previous studies [29]. In the case of the toluene-to-benzene (T/B) ratio, smaller values were obtained for the smokers group.

Figure 1 shows the ROC curves for the compounds evaluated. Xylenes gave the lowest area under the curve (AUC) with values of 0.705 (SD=0.030, 95% confidence interval, CI: 0.646-0.764, *p*<0.001) and 0.725 (SD=0.029, CI: 0.667-0.782, *p*<0.001) for o-xylene and m-, p-xylene, respectively. Toluene gave an AUC=0.753 (SD=0.028, CI: 0.698-0.808, *p*<0.001). AUC for benzene was 0.923 (SD=0.017, CI: 0.891-0.956, *p*<0.001). DMF gave an AUC=0.982 (SD=0.007, CI: 0.969-0.995, *p*<0.001). T/B ratio has an AUC=0.921 (SD=0.015, CI: 0.891-0.951, *p*<0.001). The curve for T/B ratio in Figure 1 appears below the 0.5 diagonal line (chance level) because T/B values are lower for smokers than for non-smokers.

The results of multivariate logistic regression analysis (Table 2) indicate that 2,5-dimethylfuran is the only factor independently associated with smoking status. This analyte was detected in 29 breath samples from non-smokers (14.3%), with a maximum detected level of 0.21 ppbv. A cut-off value of 0.016 ppbv (sensibility=0.965, specificity=0.896) was determined from the corresponding ROC curve using the minimum square distance method.

Breath samples from 64 patients attending two different medical practices were also analyzed and the results were compared to their self-reporting smoking status. 39 (60.1%) of the patients self-reported as being smokers and 25 (39.1%) as non-smokers. 2,5-dimethylfuran confirmed the smoking status of the 39 smokers and showed that 2 non-smokers patients (8%) had levels of 2,5-dimethylfuran above the cut-off limit.

Of the 377 subjects in this study, 36 admitted to being cannabis consumers. 29 of these (80.6%) reported that they consume cannabis mixed with tobacco, which agrees with a 2017 Global Survey that reported that 90% (80% in Spain) of European cannabis consumers smoke cannabis mixed with tobacco [40]. 2,5-dimethylfuran was detected, always above the cut-off limit, in 28 subjects from this group (96.6%). Of the 7 people that reported that they consume cannabis without mixing it with tobacco, 6 subjects yielded non-detectable levels of 2,5-dimethylfuran and one gave a level below the cut-off limit at 0.014 ppbv.

DISCUSSION

Chambers et al. [41] used the data from the 2003–2004 National Health and Nutrition Examination Survey (NHANES) to evaluate the blood levels of different VOCs and found that cigarette smoking is a primary source of benzene and toluene and an important source of xylene exposure, which was also confirmed with the data of the NHANES for the 2005-2006 period [42]. We included the T/B ratio in the present study as this ratio has been used in ambient air quality studies for estimating the ageing of air masses resulting from photochemical pollution and for characterizing the distance from vehicular emission sources, since the main anthropogenic source of VOCs in Western countries is road traffic and this ratio increases with increasing traffic volume [43,44].

Near roadsides and in urban backgrounds, typical T/B ratios are around 2.5-3, with higher ratios as the traffic volume increases or where there is the presence of industrial emissions. We have analyzed T/B ratios in 40 air samples obtained in the environments where breath samples were taken and the ratios obtained (3.348 [2.040-4.912]) are in accordance with other studies [43,44]. In the case of cigarette smoke, different studies have indicated that the T/B ratio is relatively constant, in the 1.2-2.1 range, and without significant differences for different types of cigarettes [33-35]. Although both compounds increase their levels in cigarette smoke, this increase is about 1.5 times greater in the case of benzene than of toluene [42], which leads to a decrease in the T/B ratio in cigarette smoke. The results obtained in the present study indicates that the T/B ratio for the smoker group (1.908 [1.503-2.643]) agrees with conventional emissions in cigarette smoke. For non-smokers, a significantly higher exhaled T/B ratio was found (6.965 [4.357-11.478], $p<0.001$) (Figure 2). These results confirm that the main exposure source for toluene and benzene in the smoker group is cigarette smoke.

The univariate analysis results (Table 1) show that all VOCs evaluated and the T/B ratio gave significant differences between smokers and non-smokers, which agrees with previous studies that have demonstrated that active smoking increases the levels of different VOCs related to tobacco smoke in exhaled breath [23-30], and have suggested that this matrix can be used for the assessment of smoking status. Although these results indicate that all the target VOCs may be able to assess smoking status, some studies have indicated that 2,5-dimethylfuran plays a dominant role in distinguishing between smokers and non-smokers [25,27,29,30].

A previous study showed that xylenes and toluene seem only to be adequate in the case of heavy smokers and after short-term exposure (maximum 30-45 minutes after smoking); benzene was useful for medium and heavy smokers, and for as long as 12-13 h after smoking for heavy smokers and up to 2 h for light smokers; whereas 2,5-dimethylfuran was effective for long- and short-term exposure (up to 48 after smoking) and for light and heavy consumption [29]. This study found a positive, although weak, significant correlation between the daily number of cigarettes smoked and breath levels detected. It also confirmed that breath levels are time dependent and fall rapidly after smoking. In general, it was found that breath levels depended on a combination of two parameters: time span and cigarette consumption, although time span after smoking is the most significant.

In the present study, we were focused on determining the diagnostic capacity of previously proposed VOCs without differentiating between light- and heavy-consumers and time-span, and the only limitation was of having had a minimum consumption of one cigarette per day. For this reason, the ROC curves have been determined (Figure

1) since the AUC of the ROC curve is widely recognized as the measure of a diagnostic test's discriminatory power [45]. The results obtained confirm that, according to the arbitrary classification guidelines based on a suggestion by Swets [46], xylenes and toluene have rather low accuracy (AUC<0.75), and are therefore not useful for the accurate determination of smoking status. Benzene (AUC=0.923, CI: 0.891-0.956) and T/B ratio (AUC=0.921, CI: 0.891-0.951) have good accuracy and present potential utility as a diagnostic test in some conditions, as for example short expanded time since smoking. Our results confirm that the compound with the highest discriminant capacity is 2,5-dimethylfuran with AUC=0.982 (CI: 0.969-0.995), a value that indicates excellent discriminatory ability.

The results obtained in the multivariate logistic regression analysis (Table 2) indicate that 2,5-dimethylfuran is practically the only factor that needs to be used to determine the smoking habit. The high odds ratio obtained for 2,5-dimethylfuran can be explained by the fact that this compound was not detected in 174 breath samples (85.7%) from non-smokers, and therefore the differences between the two groups are close to perfect.

2,5-dimethylfuran was detected in 29 breath samples from non-smokers (14.3%) with a median value of 0.048 [0.015-0.106] ppbv and a maximum detected level of 0.210 ppbv. It was found that in all these cases, they had drunk a cup of coffee less than 1 h before taking the sample. It has been reported that 2,5-dimethylfuran can be released by roasted coffee beans [31,32], which is due to roast defects that result in thermal degradation of D-glucose and sugar polymers [32]. To assess the effect of coffee drinking, a group of five non-smoker volunteers were asked to perform breath analysis at different times: just before drinking a coffee and at three different times afterwards (15 min, 1h, and 3h later). In no case was 2,5-dimethylfuran at the first sampling time, before drinking coffee, but the compound was detected after drinking a coffee, with a maximum level of 0.226 ppbv after 10 min. The level of 2,5-dimethylfuran decreased over time and was never detected at 3h. This indicates that coffee drinking in the last 3 h should be a requirement for a perfect discriminant detection of smoking status using breath levels of 2,5-dimethylfuran. However, we have calculated the cut-off value for 2,5-dimethylfuran in breath from the ROC curve, 0.016 ppbv (sensibility=0.965, specificity=0.896), in the case that coffee drinking is not restricted before the analysis. The results from the ten social smokers not included in the statistical evaluation showed 2,5-dimethylfuran values ranging from non-detected (n=3) to 0.050 ppbv, and 40% of them gave 2,5-dimethylfuran levels above the indicated cut-off limit.

In the case of cannabis consumers, 2,5-dimethylfuran was only detected in breath samples from those people that indicated that they mixed cannabis with tobacco and was either not detected or was below the cut-off limit for people that reported smoking cannabis alone, suggesting the value of this compound as a biomarker for tobacco use.

Finally, breath samples from 64 patients attending two different medical practices were analyzed and the results were compared to their self-reporting smoking status. 39 of the patients recognized being smokers and the results obtained for 2,5-dimethylfuran in breath confirmed their smoking status. 25 patients reported not being smokers and the breath analysis indicated that two of them (8%) should be classified as smokers. This percentage of misclassification agrees with previously reported percentages [8-14]. After a second interview with these two people, both admitted that they were in fact smokers.

The use of breath analysis, applying GC-MS, also permitted the qualitative detection of a significant presence of some VOCs that can be related to the ingestion of gums and candies or the use of toothpaste with flavors such as menthol, eucalyptol, cymene and limonene, which are present in these products. It is interesting to note that some of these compounds were found to be present in large concentrations in the breath samples of the two people who tried to hide their smoking status.

CONCLUSIONS

The results obtained confirm that 2,5-dimethylfuran is the VOC with the highest discriminant capacity for smoking status in breath analysis. This compound was the only compound tested that was able to detect smoking status in people smoking less than 1 cigarette/day and with a time-window of more than 24 h since last smoking. Benzene and T/B ratio has good accuracy for assessing the smoking status but these two parameters are not able to detect social smokers or time-windows of more than 12 h.

Despite urinary biomarkers such as NNAL being the best choice to accurately confirm smoking status, the analysis of a breath biomarker such as 2,5-dimethylfuran serves as a simple and quick check. The use of breath analysis presents many advantages over conventional blood and urine test as, in addition to its simplicity, the methodology is not invasive or embarrassing and is well accepted by patients attending clinical practices.

A limitation of our study is that we only included regular tobacco smoke exposure resulting from combustible cigarette use and so we do not know whether or not our results can be extrapolated to the use of other tobacco products.

Conflict of Interest

None declared

Acknowledgements

This study has been supported by the University of Girona (MPCUdG2016/100). The authors would like to thank A. Hughes for his help in the writing and proof reading of the manuscript.

REFERENCES

1. Das SH. Harmful health effects of cigarette smoking. *Mol. Cell. Biochem*. 2003; 253: 159-165
2. Saha SP, Bhalla DK, Whayne Jr TH, Gairola CG. Cigarette smoke and adverse health effects: an overview of research trends and future needs. *Int. J. Angiol.* 2007; 16: 77-83
3. WHO. *Assessment of the economic costs of smoking*. WHO Economics of Tobacco Toolkit, Geneva, Switzerland. 2011
4. Goodchild M, Nargis N, Tursan d'Espaignet E. Global economic cost of smoking-attributable diseases. *Tob. Control*. 2017; 0: 1-7. doi: 10.1136/tobaccocontrol-2016-053305
5. Buettgens M, Garrett B, Holahan J. American under the Affordable Care Act, Urban Institute, Robert Wood Johnson Foundation. 2010 (http://www.rwjf.org/en/library/research/2010/12/america-under-the-affordable-care-act.html) (visited 29/06/2017)
6. Kaplan CM, Graetz I, Waters TM. Most exchange plans charge lower tobacco surcharges than allowed, but many tobacco users lack affordable coverage. *Health Affairs*. 2014; 33: 1466-1473
7. Liber AC, Drope JM, Graetz I, Waters TM, Kaplan CM. Tobacco surcharges on 2015 health insurance plans sold in federally facilitated marketplaces: variation by age and geography and implication for health equity. *Am. J. Public Health*. 2015; 105, S5: S696-S698
8. Connor Gorber S, Schofield-Hurwitz S, Hardt J, Levasseur G, Tremblay M. The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status. *Nicotine Tob. Res.* 2009; 11: 12-24
9. Wagenknecht LE, Burke GL, Perkns LL, Haley NJ, Fnedman GD. Misclassification of Smoking Status in the CARDIA Study: A Comparison of Self-report with Serum Cotinine Levels. *Am. J. Public Health*. 1992; 82: 33-36
10. Coultas DB, Howard CA, Peake GT, Skipper BJ, Samet JM. Discrepancies between Self-reported and Validated Cigarette Smoking in a Community Survey of New Mexico Hispanics. *Am. Rev. Respir. Dis*. 1998; 137:810-814
11. Caraballo RS, Giovino GA, Pechacek TF, Mowery PD. Factors associated with discrepancies between self-reports on cigarette smoking and measured serum

cotinine levels among persons aged 17 years or older: Third National Health and Nutrition Examination Survey, 1988-1994. *Am J Epidemiol*. 2001; 153: 807-14

12. Caraballo RS, Giovino GA, Pechacek TF. Self-reported cigarette smoling vs. serum cotinine among U.S. adolescents. *Nicotine Tob. Res.* 2004; 6: 19-25
13. Vartiainen E, Seppälä T, Lillsunde P, Puska P. Validation of self reported smoking by serum cotinine measurement in a community-based study. *J. Epidemiol. Commun. H*. 2002; 56: 167–170
14. Wong SL, Shields M, Leatherdale S, Malaison E, Hammond D. Assessment of validity of self-reported smoking status. *Health Rep.* 2012; 23: 47-53
15. Florescu A, Ferrence R, Einarson T, Selby P, Soldin O, Koren G. Methods for quantification of exposure to cigarette smoking and environmental tobacco smoke: focus on development toxicology. *Ther. Drug Monit.* 2013; 31: 14-30
16. WHO Study Group on Tobacco Product Regulation. The specific basis of tobacco product regulation, WHO (ed), WHO Technical Report Series nº 945, p. 112, Geneva, Switzerland. 2007
17. Mattes W, Yang X, Orr MS, Richter P, Mendrick DL. *Biomarkers of Tobacco Exposure*, in: Advances in clinical chemistry, Vol. 67. Makowski GS (ed). Burlington (MA): Academic Press. pp 1-45. 2014
18. Ogburn PL, Hurt RD, Croghan IT, Schroeder DR, Ramin KD, Offord KP, Moyer PM. Nicotine patch use in pregnant smokers: nicotine and cotinine levels and fetal effects. *Am. J. Obstet. Gynecol.* 1999; 181: 736-743
19. Sällsten G, Thorén J, Barregard L, Schütz A, Skarping G. Long-term use of nicotine chewing gum and mercury exposure from dental amalgam fillings. *J. Dent. Res.* 1996; 75: 594-598
20. Yuan JM, Butler LM, Stepanov I, Hecht SS. Urinary tobacco smoke-constituent biomarkers for assessing risk of lung cancer. *Cancer Res.* 2014; 74: 401-411
21. Sandberg AS, Sköld CM, Grunewald J, Eklun A, Wheelock AM. Assessing recent smoking status by measuring exhaled carbon monoxide levels. *PLoS ONE*. 2011; 6: e28864. doi:10.1371/journal.pone.0028864
22. Borgerding M, Klusb H. Analysis of complex mixtures - cigarette smoke. *Exp. Toxicol. Pathol.* 2005; 57: 43-73
23. Perbellini I, Faccini GB, Pasini F, et al. Environmental and occupational exposure to benzene by analysis of breath and blood. *Br. J. Ind. Med.* 1988; 45: 345–352
24. Brugnone F, Perbellini I, Faccini GB, et al. Breath and blood levels of benzene, toluene, cumene, and styrene in non-occupational exposure. *Int. Arch. Occ. Env. Health.* 1989; 61: 303–311

25. Gordon SM. Identification of exposure markers in smokers' breath. *J. Chromatogr.* 1990; 511: 291–302

26. Jo WK, Pack KW. Utilization of breath analysis for exposure estimates of benzene associated with active smoking. *Environ. Res.* 2000; A83: 180–187

27. Gordon SM, Wallace LA, Brinkman MC, Callahan PJ, Kenny DV. Volatile organic compounds as breath biomarkers for active and passive smoking. *Environ. Health Perspect.* 2002; 110: 689–698

28. Van Berkel JJBN, Dallinga JW, Möller Gm, et al. Development of accurate classification method based on the analysis of volatile organic compounds from human exhaled air. *J. Chromatogr. B*. 2008; 861: 101–107

29. Alonso M, Castellanos M, Sanchez JM. Evaluation of potential breath biomarkers for active smoking: assessment of smoking habits. *Anal. Bioanal. Chem.* 2010; 396: 2987-2995

30. Gaida A, Holz O, Nell C, et al. A dual center study to compare breath volatile organic compounds from smokers and non-smokers with and without COPD. *J. Breath Res.* 2016; 10: 026006. doi:10.1088/1752-7155/10/2/026006

31. Mondello L, Costa R, Tranchida PQ, et al. Reliable characterization of coffee bean aroma profiles by automated headspace solid phase microextraction-gas chromatography-mass spectrometry with the support of a dual-filter mass spectra library. *J. Sep. Sci.* 2005; 28: 1101–1109.

32. Yang N, Liu C, Liu X, Degn TK, Munchow M, Fisk I. Determination of volatile marker compounds of common coffee roast defects. *Food Chem.* 2016; 211: 206-214

33. Charles SM, Batterman SA, Jia C. Composition and emissions of VOCs in main- and side-stream smoke of research cigarettes. *Atm. Environ.* 2007; 41: 5371-5384

34. Polzin GM, Kosa-Maines RE, Ashley DL, Watson CH. Analysis of volatile organic compounds in mainstream cigarette smoke. *Environ. Sci. Technol.* 2007; 41: 1297-1302

35. Moldoveanu S, Coleman W, Wilkins J. Determination of benzene and toluene in exhaled cigarette smoke. Beiträge zur Tabakforschung / Contributions to Tobacco Research International. 2008; 23: 107-114

36. U.S. Department of Health and Human Services. How tobacco smoke causes disease: the biology and behavioral basis for smoking-attributable disease: a report of the Surgeon General. Center for Disease Control and Prevention,

National Center for Chronic Disease Prevention and Health Promotion, Office of Smoking and Health, Atlanta (GA). 2010

37. Sanchez JM, Sacks RD. On-line multibed sorption trap and injector for the GC analysis of organic vapors in large-volume air samples. *Anal. Chem.* 2003; 75: 978-985

38. Alonso M, Castellanos M, Martin J, Sanchez JM. Capillary thermal desorption unit for near real-time analysis of VOCs at sub-trace levels. Application to the analysis of environmental air contamination and breath samples. *J. Chromatogr. B*. 2009; 877: 1472-78.

39. Alonso M, Godayol A, Antico E, Sanchez JM. Assessment of environmental tobacco smoke contamination in public premises: significance of 2,5-dimethylfuran as an effective marker. *Environ. Sci. Technol.* 2010; 44; 8289-8294

40. Winstock A, Barrat M, Ferris J, Maier L. Global Drug Survey 2017. (https://www.globaldrugsurvey.com/wp-content/themes/globaldrugsurvey/results/GDS2017_key-findings-report_final.pdf) (visited 16/01/2018)

41. Chambers DM, Ocariz JM, McGuirk MF, Bloount BC. Impact of cigarette smoking on Volatile Organic Compound (VOC) blood levels in the U.S. population: NHANES 2003-2004. *Env. Int.* 2011; 37: 1321-1328

42. Jain RB. Selected volatile organic compounds as biomarkers for exposure to tobacco smoke. *Biomarkers*. 2016; 21: 342-346

43. Gelencsér A, Siszler K, Hlavay J. Toluene-benzene concentration ration as a tool for characterizing the distance from vehicular emission sources. *Environ. Sci. Technol.* 1997; 31: 2869-2872

44. Working Group on Benzene. Position Paper Benzene. Commission of European Communities. Council Directive on Ambient Air Quality Assessment and Management. 1998 (http://ec.europa.eu/environment/air/pdf/ppbenzene.pdf) (visited 29/07/2017)

45. Faraggi D, Reiser B. Estimation of the area under the ROC curve. *Stat. Med.* 2002; 21: 3093-3106

46. Swets JA. Measuring the accuracy of diagnostic systems. *Science*. 1988; 240: 1285-1293

Table 1. Median and quartiles (25 and 75%) for the analyte concentrations and T/B ratios in the breath of smokers and non-smokers

	Smokers (n=174)	Non-smokers (n=203)	*p*-value
o-xylene	0.366 [0.248-0.641]	0.244 [0.172-0.329]	**< 0.001**
m-, p-xylene	1.019 [0.494-1.507]	0.481 [0.284-0.836]	**< 0.001**
Toluene	5.109 [2.045-9.824]	1.653 [0.853-3.336]	**< 0.001**
Benzene	2.530 [0.798-6.095]	0.283 [0.116-0.419]	**< 0.001**
2,5-dimethylfuran	0.303 [0.085-0.882]	nd [nd-nd] [(a)]	**< 0.001**
T/B ratio	1.908 [1.503-2.643]	6.965 [4.357-11.478]	**< 0.001**

Concentrations are expressed in parts per billion by volume (ppbv)

[(a)] nd: not detected (a value of 0.0 was used for statistical analysis)

Table 2. Logistic regression analysis of smoking status

	Odds ratio	Confidence Interval, 95%	*p*-value
o-xylene	0.286	0.051-1.590	0.153
m-, p-xylene	3.381	0.579-19.753	0.176
Toluene	1.057	0.337-3.321	0.924
Benzene	0.007	0.000-23.289	0.232
2,5-dimethylfuran	$2.6 \cdot 10^{65}$	$3.0 \cdot 10^{35}$-$2.1 \cdot 10^{95}$	**< 0.001**
T/B ratio	0.858	0.597-1.233	0.407
Age	0.977	0.905-1.055	0.559
Sex	4.341	0.679-27.751	0.121

Figure captions

Figure 1. Receiver operating characteristic curves for target analytes and T/B ratio. The sensitivity on the ordinate represents the true positive rate whereas 1-specificity represents the false positive rate, where the specificity is the true negative rate.

Figure 2. Box-plots for the data obtained for the different compounds and parameters evaluated

Figure 1

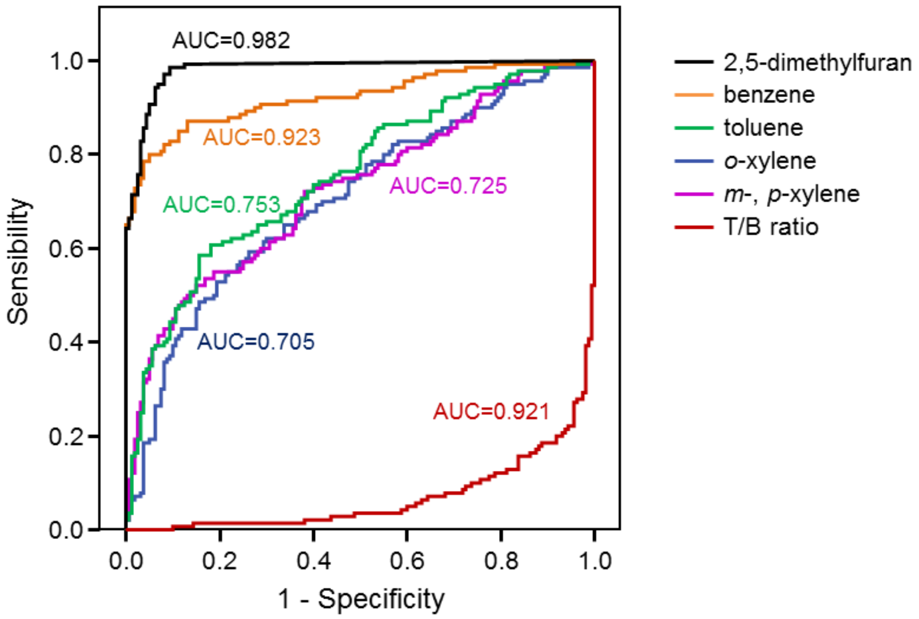
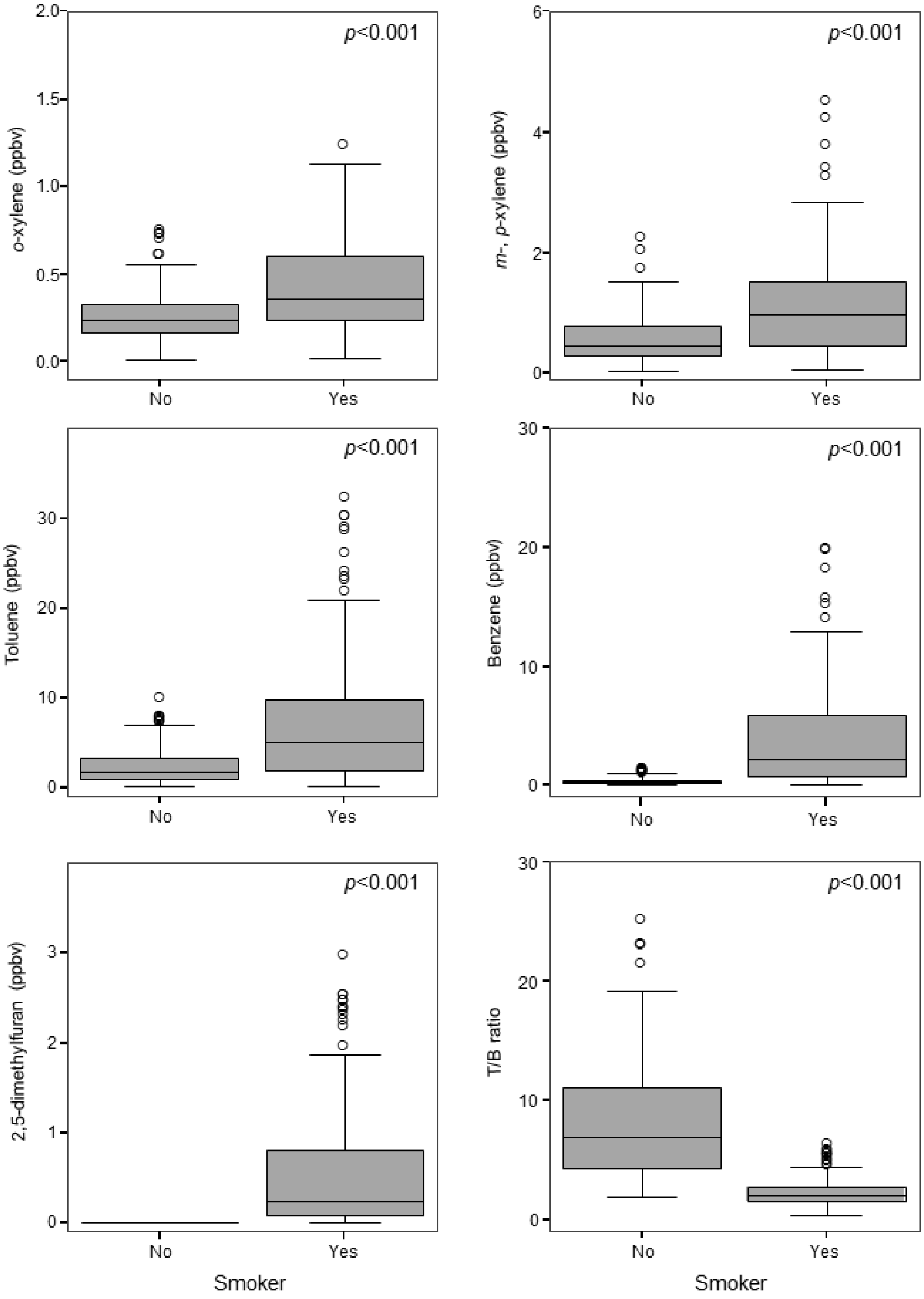

AUC=0.982
AUC=0.923
AUC=0.753
AUC=0.725
AUC=0.705
AUC=0.921
Sensibility
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1 - Specificity
2,5-dimethylfuran
benzene
toluene
o-xylene
m-, *p*-xylene
T/B ratio

Figure 2.



o-xylene (ppbv)
2.0
1.5
1.0
0.5
0.0
p<0.001
No
Yes
m-, p-xylene (ppbv)
6
4
2
0
p<0.001
No
Yes
Toluene (ppbv)
30
20
10
0
p<0.001
No
Yes
Benzene (ppbv)
30
20
10
0
p<0.001
No
Yes
2,5-dimethylfuran (ppbv)
3
2
1
0
p<0.001
No
Yes
Smoker
T/B ratio
30
20
10
0
p<0.001
No
Yes
Smoker