HHS Public Access

Author manuscript

Environ Sci Technol. Author manuscript; available in PMC 2018 January 03.

Published in final edited form as:

Environ Sci Technol. 2018 January 02; 52(1): 308–316. doi:10.1021/acs.est.7b05128.

Distinguishing petroleum (crude oil and fuel) from smoke exposure within populations based on the relative levels of benzene, toluene, ethylbenzene, and xylenes (BTEX), styrene and 2,5-dimethylfuran by pattern recognition using artificial neural networks

D.M. Chambers[*], **C.M. Reese**, **L.G. Thornburg**, **E. Sanchez**, **J.P. Rafson**, **B.C. Blount**, **J.R.E. Ruhl III**, and **V.R. De Jesús**

Tobacco and Volatiles Branch, Division of Laboratory Sciences, National Center for Environmental Health, US Centers for Disease Control and Prevention

Abstract

Studies of human exposure to petroleum (crude oil and fuel) often involve monitoring volatile monoaromatic compounds because of their toxicity and prevalence. Monoaromatic compounds such as benzene, toluene, ethylbenzene, and xylenes (BTEX) associated with these sources have been well studied and have established reference concentrations (RfC) and reference doses (RfD). However, BTEX exposure levels for the general population are primarily from tobacco smoke, where smokers have blood levels up to 8 times higher on average than nonsmokers. Therefore, in assessing petroleum exposure, it is essential to identify exposure to tobacco smoke as well as other types of smoke exposure (e.g., cannabis, wood) because many smoke volatile organic compounds are also found in petroleum products such as crude oil, and fuel. This work describes a method using partition theory and artificial neural network (ANN) pattern recognition to accurately categorize exposure source based on BTEX and 2,5-dimethylfuran blood levels. For this evaluation three categories were created and include crude oil/fuel, other/nonsmoker, and smoker. A method for using surrogate signatures (i.e., relative VOC levels derived from the source material) to train the ANN was investigated where blood levels among cigarette smokers from the National Health and Nutrition Examination Survey (NHANES) were compared with signatures derived from machine-generated cigarette smoke. Use of surrogate signatures derived from machine-generated cigarette smoke did provide a sufficient means with which to train the ANN. As a result, surrogate signatures were used for assessing crude oil/fuel exposure because there is limited blood level data on individuals exposed to either crude oil or fuel. Classification agreement between using an ANN model trained with relative VOC levels and using the 2,5-dimethylfuran smoking biomarker cutpoint blood level of 0.014 ng/mL was up to 99.8 % for nonsmokers and 100.0% for smokers. For the NHANES 2007–08 data, the ANN model using a probability cutpoint above 0.5 assigned 7 samples out of 1998 (0.35%) to the crude oil/fuel signature category. For the

[*]Corresponding author: 4770 Buford Hwy., NE, Mail Stop F-47, Atlanta, GA 30341, mzz7@cdc.gov.

Disclaimer

The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention. Use of trade names is for identification only and does not imply endorsement.

NHANES 2013–14 data, 12 out of 2906 samples (0.41%) were assigned to the crude oil/fuel signature category. This approach using ANN makes it possible to quickly identify individuals with blood levels consistent with a crude oil/fuel surrogate among thousands of results while minimizing confounding from smoke. Use of an ANN fixed algorithm makes it possible to objectively compare across populations eliminating classification inconsistency that can result from relying on visual evaluation.

1. Introduction

As oil and natural gas extraction capacity grows to meet increased demand for petroleum-based products, the potential for exposure to volatile organic compounds (VOCs) from these sources increases. Exposure to VOCs from these particular sources is important to identify and minimize because they include toxicants such as benzene, toluene, ethylbenzene, and xylenes (BTEX) that can cause cancer, neurological damage and impairment, and pulmonary and cardiovascular disease (Grandjean and Landrigan 2014; Korte et al. 2000; Maltoni et al. 1985; Mogel et al. 2011; Xu et al. 2009). BTEX concentrations are commonly measured in environmental and biomonitoring samples to assess exposure levels associated with petroleum sources because they are (1) prominent components in petroleum, (2) carried over from crude oil into consumer products or fortified in fuels for their anti-knock properties, (3) chemically stable, and (4) persist as intact compounds in the environment and in the body.

Although BTEX are quantifiable in breath, blood, urine and as urinary metabolites, blood levels are more sensitive (low pg/mL), selective, and can be more precisely quantified (Blount et al. 2006; Chambers et al. 2006). In general, because of their relatively nonpolar character, VOCs such as BTEX partition at higher concentrations in blood than in urine. Additionally, if the VOC is not reactive, it will not efficiently form a metabolite. However, in the case of reactive VOCs such as acrolein, acrylonitrile, 1,3-butadiene, and isoprene, VOC concentrations should be measured using the metabolite as these compounds are not stable in blood on the order of days, even if stored at refrigerator temperatures. Even though reactive VOCs can be captured in breath in real time, quantitative breath measurements of VOCs are challenging to standardize because VOCs localize in different regions of the lungs (Anderson et al. 2003; Kim et al. 2012). In addition, mixing of lung air containing VOCs with inhaled air can cause variations in measured levels that are dependent on breathing technique. Because lung retention has been shown to be related to blood/air partition constant (Jakubowski and Czerczak 2009), BTEX levels are between 7 to 35 times higher in blood than in lung air. These variations in breath analysis makes this measurement more suitable for noninvasive semi-quantitative or qualitative analyses.

A reoccurring impediment to petroleum exposure assessment using biomarkers such as BTEX has been confounding by exposure to tobacco smoke VOCs because BTEX are formed from tobacco combustion. In fact, tobacco smoke contains ppm levels(Pazo et al. 2016) of these and other petroleum biomarkers as both are derived from the decomposition of biomass. Among cigarette smokers, blood VOC levels can reach the low ng/mL range, whereas nonsmokers are typically below the pg/mL range (Chambers et al. 2011).

Therefore, if including smokers in biomonitoring studies investigating petroleum exposure, smokers need to be accurately categorized because of their potential to confound the assessment.

To advance more accurate and consistent exposure assessment, we describe a biomonitoring approach that identifies VOC exposure specific to petroleum through pattern recognition using artificial neural network modeling. As a way to minimize confounding seen by other significant VOC exposure sources such as tobacco smoke, we investigated the use of relative combustion biomarker levels of BTEX and styrene (BTEXS), and 2,5-dimethylfuran. Blood levels for styrene are included in this work because styrene is a BTEX concomitant measurable in smoke and petroleum (Chambers et al. 2006; Chambers et al. 2011). 2,5-Dimethylfuran is included because it is a smoke biomarker that exists at levels comparable to BTEX and styrene (Ashley et al. 1996; Gordon SM 1990). In this work, we demonstrate that relative blood BTEXS levels among smokers and cigarette tobacco smoke remain consistent over a range of cigarette brand varieties and smoking techniques. This approach, based on the similarity of BTEXS physical properties and on partition theory, is used to identify individuals with relative levels of these VOCs in their blood that are consistent with those deduced from concentrations associated with petroleum (e.g., crude oil and fuel). Although VOC blood levels vary with source concentration, relative proportions can remain consistent and unique to a particular source especially for VOCs with similar chemical properties, such as the BTEXS.

2. Experimental

2.1. Population Data

Population data were taken from the 2007–08 and 2013–14 National Health and Nutrition Examination Survey (NHANES) provided by the National Center for Health Statistics (NCHS) (Centers for Disease Control and Prevention (CDC) 2007–08 and 2013–14). Collection of blood is a priority exam at the NHANES mobile examination center, especially if participants have been fasting for at least 9 hours. As a result, the mean collection time from check-in is approximately 40 min. Laboratory measurements were performed by CDC's Division of Laboratory Sciences (DLS) at the National Center for Environmental Health (NCEH). The VOC blood levels were quantified by equilibrium headspace solid phase microextraction (SPME)/gas chromatography (GC)/mass spectrometry (MS) of hermetically collected participant blood samples collected at an NHANES mobile examination center (MEC) (Centers for Disease Control and Prevention (CDC) 2007–08 and 2013–14). Blood from 3415 participants for NHANES 2007–08 (SAS export file: VOCWB_E.XPT) and 3489 for NHANES 2013–14 (SAS export file: VOCWB_H.XPT) was analyzed, where both cycles were a half subsample of 12 years of age and older. When available, blood levels for benzene, toluene, ethylbenzene, m/p-xylene, o-xylene, styrene, and 2,5-dimethylfuran were used in this study. Specifically, styrene was not available in the NHANES 2013–14, but was available in the NHANES 2007–08 and thus was used in pattern recognition modeling for other/nonsmoker and smoker. Participants were excluded if any of the needed VOCs were missing. Survey data for the 2007–08 NHANES cycle on recent VOC exposure was included in SAS export file, VOCWB_E.XPT, and recent tobacco

use were collected at the mobile examination center (SAS export file: SMQRTU_E.XPT) on the day of the health exam. Corresponding survey data for the 2013–14 NHANES cycle on recent VOC exposure was reported in SAS export file VTQ_H.XPT and recent tobacco use in file SMQRTU_H.XPT. Specific questionnaire data regarding recent VOC exposure and tobacco use for evaluating ANN model results are specified in the Discussion section.

2.2 Estimation of VOC Surrogate Signatures from Source Concentration

Cigarette tobacco smoke BTEXS and 2,5-dimethylfuran levels used in this study were quantified using Tedlar bag collection of machine smoked cigarettes followed by equilibrium headspace SPME/GC/MS analysis as described elsewhere (Sampson et al. 2014). These cigarette smoke VOC levels were generated using the ISO (3308:2012) and Canadian Intense (CI) protocols from 50 U.S. brand varieties of cigarettes from brand families that comprise 78% of the market share at the time of the sampling (Pazo et al. 2016). Surrogate estimates of blood BTEXS and 2,5-dimethylfuran levels were produced by multiplying these smoke levels by their respective blood/air partition constant, $K_{blood/air}$, and normalizing the levels relative to toluene, a high detection rate analyte in blood. Although, m/p-xylene blood levels had a slightly higher detection rate than toluene, toluene was used because available crude oil and fuel BTEX data combined m/p-xylene with o-xylene where o-xylene had a lower detection rate than toluene. The corresponding $K_{blood/air}$ values used for this adjustment were taken from Meulenberg and Vijverberg (Meulenberg and Vijverberg 2000), where benzene is 7.37, toluene is 15.11, ethylbenzene is 28.2, styrene is 55.6, total xylene is 36.13 (the average of 33.2 for m-xylene, 38.9 for p-xylene, and 36.3 for o-xylene), and m/p-xylene is 36.05 (the average $K_{blood/air}$ for m-xylene and p-xylene). A measured value of $K_{blood/air}$ for 2,5-dimethylfuran was not available, but was estimated to be 8.5 using previous published regression data (Kramer et al. 2016). The use of these surrogate VOC signatures (i.e., relative VOC levels derived from the source material) is compared with measured blood VOC signatures (i.e., relative VOC levels in blood) in training of the artificial neural network pattern recognition model for data classification.

The surrogate blood VOC signatures for crude oil and fuel exposure used for training the ANN were estimated from crude oil and fuel source signatures from the Environment Canada Environmental Technology Centre Oil Properties Database (Environment Canada Environmental Technology Centre 2001). This database reported data for benzene, toluene, ethylbenzene and total xylene, but not styrene. Unweathered crude oil selected were representative of different regions of the world and fuels were representative of different types of fuels. Crude oils included Cusiana (Columbia), Dos Cuadras (California), Boscan (Venezuela), Arabian Medium, Alaska North Slope (2002), Arabian Light (2000), Chayvo #6 (Sakhalin), Amauligak (Canada), Carpinteria (California), Maya (Mexico), Maya (1997), Isthmus (Mexico), Cano Limon (Columbia), Vasconia (Columbia), BCF 24 (Venezuela), West Texas Intermediate, West Texas (2000), West Texas Sour, Cold Lake Blend (Alaska), Mississippi Canyon Block 194, Terra Nova (Canada), Eugene Island Block 32 (Louisiana), South Louisiana (2001), Ekofisk (Norway). Fuels included Aviation Gasoline 100LL, Diesel Fuel Oil (2002), Diesel Fuel Oil (Alaska), Diesel Fuel Oil (Southern U.S.A., 1994), Diesel Fuel Oil (Southern U.S.A., 1997), Fuel Oil No. 5 (2000), High Viscosity Fuel Oil, Jet A/Jet A-1, Jet B (Alaska). The automobile gasoline BTEX signature used was from the Total

Petroleum Hydrocarbon Criteria Working Group (Potter et al. 1998). These surrogate blood signatures were created using the same procedure as described for producing the surrogate blood VOC signatures for cigarette smoke exposure.

2.3 Statistical Analysis

JMP 12.0 was used for all the statistical analyses. Central tendency for blood VOC levels is expressed as geometric mean because the data followed a log normal distribution. For the artificial neural network (ANN) analyses, model parameters were fit via penalized maximum likelihood maximization (Gotwalt 2011). Separate log-likelihoods were computed for each response, and the overall log-likelihood for all the responses was taken as the sum of the log-likelihoods of the individual responses. The model used one hidden layer with 15 hidden nodes each using a hyperbolic tangent (TanH) activation function and was validated using a holdback fraction of 0.33. The optimal number of hidden nodes was determined as the minimal number of nodes that yielded, on average, the best classification accuracy in the training set with regard to the surrogate crude oil/fuel signatures. The output layer had three output nodes, corresponding to smoker, to crude oil/fuel, and to neither smoker nor crude oil/fuel, which is referred to as other/smoker. Input variables for the characterization of smoker and other/nonsmoker (not exposed to either crude oil/fuel or smoke) included blood levels for BTEXS and 2,5-dimethylfuran where smoking status was designated by 2,5-dimethylfuran blood levels greater than or equal to 0.014 ng/mL. Input variables for the characterization of crude oil/fuel exposure included normalized surrogate blood signatures calculated from measurements of source petroleum/fuels for BTEX where 2,5-dimethylfuran was inputted as below LOD. Models that involved categorizing only smokers and other/nonsmokers used the entire 2007–2008 NHANES data set where there were approximately 4 times more other/nonsmokers than smokers. However, for models with the crude oil category, all the categories were adjusted so that each category had similar numbers of results of approximately 500, 333 for training and 167 for validating the ANN. In this case, the training set for the other/nonsmoker category was randomly sampled to decrease its size, all smokers were used and the surrogate crude oil/fuels signatures were oversampled(Chawla et al. 2002) by duplicating their results. Output value (probability ranging from 0 to 1) cutpoint for assignment to a category is set to 0.5. Blood levels were normalized because the surrogate signature used for crude oil and fuel exposure only expresses relative level and not absolute concentration. Rows with missing values were ignored.

The robustness and appropriateness of using measured vs. surrogate signatures were assessed in an experiment involving classification of smokers and other/nonsmokers using the NHANES 2007–2008 data. The percentage of correct classification assignments was based on 2,5-dimethylfuran cut-point. Note that the NHANES 2007–2008 data had only two classifications, other/nonsmoker and smoker, any individuals that may have a signature corresponding to the surrogate crude oil/fuel signature were initially assigned to either other/nonsmoker or smoker in the training. Following this experiment, an ANN model used for classification of all three groups—other/nonsmoker, smoker, and crude oil/fuel—was constructed by combining the surrogate crude oil/fuel signatures with the other/nonsmoker and smoker signatures from the NHANES 2007–2008 data. Once this final model with the three categories was created, it was applied to the NHANES 2013–2014 data.

3. Results

Without a sufficient quantity of known crude oil or fuel blood signatures with which to train the ANN, it was necessary to generate surrogate signatures to flag possible crude oil or fuel exposed individuals at large. For this surrogate signature, blood BTEX signatures were estimated from relative levels of these chemicals in different crude oils and fuels. Estimates of relative levels of VOCs in blood were made by adjusting the relative crude oil or fuel BTEX concentrations based on blood-air partition constant, $K_{blood/air}$, assuming equilibration. This approach is first demonstrated in the following data where relative blood VOC levels of cigarette smokers are estimated from relative VOC levels measured in cigarette smoke from different cigarettes.

Shown in Fig. 1 is a comparison between absolute and relative blood VOC levels measured in cigarette smokers and relative levels measured in cigarette smoke itself. The BTEXS and 2,5-dimethylfuran blood VOC levels were from the 2007–08 NHANES and are baseline adjusted mean levels for participants who reported smoking only 5, 10, 15, 20, 30, and 40 cigarettes per day (CPD). The number of individuals (N) in these categories ranged from 14 to 91 as noted in the figure caption. Baseline adjustment of cigarette smoker blood levels was performed by subtracting the mean blood level of individuals that were classified as other/nonsmoker (N=4876). Other/nonsmoker were identified as having blood 2,5-dimethylfuran levels < 0.014 ng/mL. Levels in Fig. 1a were not normalized so that magnitude of VOC exposure could be compared. Upon normalizing Fig. 1a data to toluene, the relative standard deviation of the relative signal that existed across the CPD categories in Fig. 1a was 10.7 for m/p-xylene, 8.1 for benzene, 7.6 for ethylbenzene, 18.9 for styrene, 29.6 for o-xylene, and 9.6 for 2,5-dimethylfuran.

The surrogate blood VOC signatures estimated from cigarette smoke levels are shown in Fig. 1b. These signatures were created by first multiplying the amount of the VOC per smoked cigarette (generated using either the Canadian Intense and ISO protocols) by the corresponding $K_{blood/air}$, and then normalizing these levels with respect to toluene for the sake of comparison with blood levels in cigarette smokers. This surrogate signature is compared to a normalized composite of blood levels among all NHANES 2007–08 smokers, classified using a 2,5-dimethylfuran cutpoint level above 0.014 ng/mL.

The same procedure used to estimate blood VOC level signatures for smokers was used to generate potential blood VOC signatures for individuals with blood VOC signatures consistent with those for deduced from crude oil or fuel. In this application, VOC levels among different crude oil and fuels were obtained primarily from Environment Canada's Oil Properties Database. Petroleums selected included 25 crude oils from the United States and other large oil producing countries, and 9 fuels (Environment Canada Environmental Technology Centre 2001). Because a signature for automobile gasoline was not available in this database, the one from the Total Petroleum Hydrocarbon Working Group was used (Potter et al. 1998). In the Environment Canada's database, levels of xylene's structural isomers (m-xylene, p-xylene, and o-xylene) were combined and there were no data for styrene. Levels for 2,5-dimethylfuran, which is not a substantial compound in petroleum, were added to the signature and set to below the LOD (imputed as LOD/√2). Accordingly,

the BTEX levels for different petroleums were adjusted with the corresponding $K_{blood/air}$ and normalized relative to toluene for intercomparison and comparison with the NHANES blood signatures. Shown in Fig. 2 is a comparison of the resulting composite petroleum signature, the surrogate blood level composite signature derived from the petroleum signatures (adjusted based on blood/air partition constants and normalized relative to toluene), and an individual blood signature from a known fuel exposure (Chambers et al. 2008).

Shown in Fig. 3 are density plots characterizing ANN predictions for the NHANES 2013–14 data using a model trained and validated with NHANES 2007–08 data. Toluene, total xylenes (*m*/*p*-xylene + *o*-xylene), benzene, ethylbenzene, and 2,5-dimethylfuran levels were used as input variables in modeling. For the ANN model, the prediction variables were other/nonsmoker, smoker, and crude oil/fuel groups, where training signatures were from individuals identified as other/nonsmoker and smokers established using blood 2,5-methylfuran levels (cutpoint = 0.014 ng/mL)(Chambers et al. 2011) and surrogate crude oil and fuel signatures. The training set for the other/nonsmoker signature included crude oil/fuel signatures that had not yet been identified as so. Fitting the NHANES 2013–14 data repeatedly ten times resulted in varying assignments where the number of predicted crude oil/fuel signatures ranged from 7 to 24. Each of these 10 models properly identified a positive control blood signature from an individual with known fuel exposure presumed to be from inhalation(Chambers et al. 2008) with probabilities ranging from 0.9800 to 0.9999. Visual inspection of each sample identified as having the crude oil/fuel signature by any of these 10 models were similar to the surrogate and known exposure blood sample signatures. The first model fit of the NHANES 2007–08 data categorized 7 individuals as crude oil/fuel (0.35%), 1591 as other/nonsmoker (79.63%) and 400 as smoker (20.02%). Prediction results using the NHANES 2013–14 blood VOC data from the first model fit are graphed in Fig. 3a and consisted of 12 crude oil/fuel (0.41%), 2290 other/nonsmoker (78.80%) and 604 smoker (20.79%). Note that the crude oil/fuel probability cutpoint is 0.5, however because of smoothing by JMP, the graphical boundary for the crude oil/fuel distribution extends below 0.5. Most smokers and other/nonsmoker had a probability below 0.25 with 5 other/nonsmoker and 1 smoker between 0.25 and 0.5. Shown in Fig. 3b are corresponding probabilities for other/nonsmoker and smoker with the petroleum/fuel group identified in Fig. 3a. Among the samples categorized as either other/nonsmoker or smoker, most (i.e., 2791) had probabilities less than 0.25 or greater than 0.75, where 103 samples fell between these limits.

Compared in Fig. 4 are VOC levels categorized in the three exposure categories including crude oil/fuel, other/nonsmoker, and smoker groups for the NHANES 2007–08 data with m/p-xylene and o-xylene separated and with styrene included. Blood concentrations are expressed as geometric means because the data are lognormally distributed. Relative toluene and 2,5-dimethylfuran levels are highest among smokers. Among individuals categorized in the crude oil/fuel group had relative xylene levels significantly higher than other/nonsmoker and smoker. Benzene and styrene levels were comparable between smoker and crude oil/fuel groups.

4. Discussion

Elimination half-life for BTEXS, which have similar blood stabilities, polarities and solubilities, range from 15–30 hrs and are typically 1–2 orders of magnitude longer than alpha-phase half-lives. Therefore, blood levels can persist above detectable levels even if blood samples are collected beyond the alpha-phase half-life. Furthermore, an average systematic concentration or steady state is achieved for persistent transient exposure whose mean can be determine with sufficient sampling of the population. Individuals who are not exposed to either crude oil, fuel, or smoke have BTEXS blood levels near or below detection as these VOC sources substantially drive BTEXS levels in the general U.S. population (Chambers et al. 2011). BTEXS signatures resulting from crude oil/fuel exposure can be confounded by exposure to tobacco smoke, especially from cigarette use, as smoking prevalence in the United States is 16.4% (Jamal et al. 2015) and tobacco smoke has high levels of BTEXS (μg/cigarette) (Pazo et al. 2016). Fortunately, the blood BTEXS signature among smokers (based on relative level) is conserved despite demographic and smoking technique differences, although absolute levels vary substantially. This consistency in relative blood BTEXS level is apparent from the high correlation among the BTEXS compounds for smokers who report smoking between 5 and 40 cigarettes per day (Fig. 1a), where Pearson r ranges from 0.55 (o-xylene and benzene) to 0.96 (m/p-xylene and ethylbenzene) and are consistent with those previously reported (Chambers et al. 2011). The basis for this consistency is the conservation of relative VOC levels from mainstream cigarette smoke where these relative levels have been shown to exist within a narrow range. In a study of 50 different U.S. brand varieties generated with two different smoking machine protocols, the Pearson r values among the BTEXS ranged from 0.83 (o-xylene vs. benzene) to 0.98 (m/p-xylene vs. o-xylene) (Pazo et al. 2016). This high degree of consistency and correlation persists despite differences in influential cigarette parameters such as percent tip ventilation and number of puffs per cigarette (Pazo et al. 2016) and confirms that there is a characteristic BTEXS signature among the different brand varieties.

This consistency in relative BTEXS levels in mainstream smoke makes possible a proportional relationship between tobacco smoke and blood VOC levels that is attributed to quick equilibration of inhaled VOCs. Blood VOC levels dependent on blood-air partitioning, $K_{blood/air}$, equilibrate quickly because of relatively small pulmonary venous blood volume and high flowrate (Anderson et al. 2003; Wagner et al. 1974). This relationship is confirmed by comparing smoker blood levels with smoke VOC signatures taken from machine-generated levels that have been adjusted with the respective $K_{blood/air}$ (i.e., multiplying smoke VOC levels by their $K_{blood/air}$). During exposure to the VOCs in cigarette smoke, the proportions of VOCs in the blood are based on their affinity that corresponds to $K_{blood/air}$. Because magnitude of VOC levels can vary depending on smoking technique, that is, the manner in which a smoker smokes a cigarette, two machine smoking protocols with different smoking intensities were evaluated— the ISO and Canadian Intense (CI) (Bialous and Yach 2001; Health Canada 1999). Shown in Fig. 1b is a comparison of the normalized adjusted BTEXS and 2,5-dimethylfuran signatures produced by the ISO and CI protocols from a U.S. market study of 50 brand varieties. Although the signatures associated with these two protocols are similar, there are some minor differences that are statistically

significant. Specifically, *m*/*p*-xylene and styrene levels were lower relative to toluene for the ISO protocol than with the CI protocol. Because the blood BTEXS and 2,5-dimethylfuran signature of smokers more closely resembles that produced by the ISO protocol than the CI protocol (Fig. 1b), the ISO signature was used in ANN training comparison experiments discussed below. The similarity between VOC levels in smokers and adjusted VOC levels in smoke generated by the ISO method is a noteworthy finding, and underscores the consistency of relative BTEXS levels in cigarette smoke and among smokers.

Because of the consistency of relative blood levels of BTEXS/2,5-dimethylfuran among smokers given vastly different exposure magnitudes, we investigated whether a BTEXS/2,5-dimethylfuran signature associated with crude oil or fuel could be identified in the general population using ANN pattern recognition. To train the ANN, best practice is to use relative levels of blood VOCs of petroleum exposed individuals. The limitation in using surrogate signatures produced from source concentrations to train the ANN is that they lack magnitude information, which is needed to properly add the contribution of any baseline levels that might exist. For example, although the ANN models constructed with signatures from either confirmed smokers or the surrogates produced similar outcomes for distinguishing between confirmed smokers and other/nonsmokers, not being able to add a baseline level to surrogate signatures did cause the model to classify individuals with low-level smoke VOC exposure as other/nonsmoker as the baseline level became a more prominent component of the signature. For reference, in the comparison shown in Fig. 1b, the mean baseline level of nonsmokers was subtracted from the smoker composite signature. In reality, the smoker composite includes baseline levels as shown in Fig. 4. With this baseline level absent from the surrogate signatures, the ANN (over 10 model iterations) agreed with the classification as defined by the 2,5-dimethylfuran cutpoint for 99.5–99.9 % of nonsmokers (N=1502), but only 85.7–90.7 % of smokers (N=356). On the contrary, use of blood VOC signatures from smokers resulted in classification of 99.0–99.8 % of nonsmokers (N-1502) and 97.5–100.0% of smokers (N=356) agreement with the 2,5-dimethylfuran cutpoint (over 10 model iterations).

For ANN analyses that used known smokers, smokers were identified using the smoke biomarker 2,5-dimethylfuran with a cutpoint of 0.014 ng/mL. It was convenient to use 2,5-dimethylfuran levels to classify smokers because 2,5-dimethylfuran is measured simultaneously with the other VOCs and is a well-established smoking biomarker (Ashley et al. 1996; Gordon SM. 1990). Training and validation of the ANN involved using the NHANES 2007–08 data and excluding samples with missing data. The ANN model categorized participants who reported recent use of other combustible tobacco products [e.g., pipe (SMQ755 = 1, N = 1), cigar (SMQ 785 = 1, N = 6)] as cigarette smokers, with the exception of three cigar smokers. Two of these cigar smokers (SEQN# 46111 and 47799) had 2,5-dimethylfuran levels below the established 0.014 ng/mL cutpoint and were identified as nonsmokers in all 10 iterations of the model. VOC levels for these two samples are more consistent with those of the other/nonsmoker category having m/p-xylene below 0.1 ng/mL. The third cigar smoker (SEQN# 51320) had a blood 2,5-dimethylfuran level of 0.055 ng/mL, but was classified as an other/nonsmoker in 5 of the 10 ANN models, where the smoker probability average was 0.52 and other/nonsmoker average was 0.48. This average result, which assigns this third cigar smoker as a smoker, suggests that averaging

probabilities over multiple ANN model instances may better predict borderline signatures. Furthermore, recent marijuana use did not interfere with proper classification of tobacco smokers. Specifically, among the 3 tobacco smokers who reported recent marijuana use (DUQ220Q = 0/DUQ220U = 1), all signatures were assigned as smoker and all had blood 2,5-dimethylfuran level above the smoker biomarker cutpoint.

The petroleum blood signature was defined using a surrogate approach because of a lack of VOC data for petroleum exposed individuals. To create crude oil and fuel signatures with which to train the ANN, crude oil and fuel levels from the Oil Properties Database(Environment Canada Environmental Technology Centre 2001) and a literature source for automobile gasoline(Potter et al. 1998) were multiplied by the corresponding $K_{blood/air}$ and normalized relative to toluene as demonstrated with the cigarette smoke signatures. Despite the fact that fuel composition varies across crude oil sources, manufacturers, and the time of year (Potter et al. 1998; Wang et al. 2003), especially in terms of absolute levels, relative BTEX levels among low-end (e.g., gasoline) and middle distillates (diesel, jet, and home heating fuel) were similar to those seen in crude oil. This consistency is apparent in Fig. 2 for the adjusted and normalized BTEX signature where standard deviation error bars do not overlap. Relative total xylene levels were highest, toluene and ethylbenzene levels were moderate and benzene levels were the lowest. These adjusted crude oil and fuel signatures are consistent with previously published blood VOC data resulting from a fuel inhalation exposure subject shown in Fig. 2, which was used as a blinded positive control (Chambers et al. 2008).

Using the ANN model, 12 individuals out of 2906 had exposure patterns that placed them in the crude oil/fuel exposure category with a probability that ranged from 0.52 to 0.98 (Fig. 3a). Of the 12 crude oil/fuel categorized individuals, all were nonsmokers based on the 2,5-dimethylfuran cutpoint and questionnaire responses. There was no consistent trend with recent use of gas for cooking (VTQ241A, 4 out of 12), pumping gas or diesel (VTQ244A and VTQ281C, 7 out of 12), or time spent near smoke in the last 10 hours or less (VTQ265B, 1 out of 12), however all had reported having an attached garage (VTQ210, 12 out of 12). Association between attached garage and increased blood BTEX levels has been previously reported (Mallach et al. 2017; Symanski et al. 2009). This association is attributed to outgassing of butyl rubber tires(Chambers et al. 2006) and evaporative fuel emissions from vehicles, which have been identified as an important sources of VOC emissions in the U.S. and other industrialized nations (Liu et al. 2015; United States Environmental Protection Agency 2015; Yamada 2013).

Of the 32 individuals that reported pumping gas in the last hour, one (SEQN# 80105) was classified as crude oil/fuel and had a high m/p-xylene level of 1.140 ng/mL. It is important to note that this identification was made based on relative levels of VOCs and not magnitude because data fed into the model was normalized relative to toluene. The nonsmoker classification had 19 individuals with m/p-xylene blood levels ranging from <0.024 to 0.159 ng/mL and smoker classification included 7 individuals with m/p-xylene ranging from 0.077 to 0.207 ng/mL. All smokers had 2,5-dimethylfuran above the 0.014 ng/mL cutpoint with the exception of SEQN# 78168. This sample (# 78168) was categorized as a smoker in all 10 consecutive runs of the ANN model with categorization probability for those runs ranging

from 0.82 to 1.00. However, this individual had self-reported smoking marijuana that day (DUQ220Q = 0/DUQ220U = 1). Because ANN assignments are restricted to only the categories available in the model, exposure from a different source will be assigned to the category that most closely represents the signature. In the case of the self-reported marijuana smoker, the overall signature was assigned to the category trained on tobacco smokers. The remaining 5 individuals were not included because of missing blood level data. These data suggest that most individuals, 31 out of 32, who pumped gas within an hour before the MEC exam did not experience a level of BTEXS exposure to significantly alter their blood signature and absolute levels were commensurate with their ANN assigned category.

Although styrene levels were not available for ANN modeling, signatures for NHANES 2007–08 data are shown in Fig. 4 so as to demonstrate relative styrene blood levels, which are an important concomitent in crude oil/fuel and smoke blood signatures. Geometric means are plotted because VOC distributions are log normal. The presentation of the smoker category signature differs slightly from that shown in Figs. 1a and 1b, which were presented as arithmetic means and adjusted by subtracting the mean baseline level taken from nonsmokers. VOC variability was smallest among smokers where relative standard deviation ranged from 61 to 85% for the BTEXS and 2,5-dimethylfuran. RSDs for monoaromatic compounds among those categorized as other/nonsmoker ranged from 57% to 528%, where styrene (RSD = 57%) had the lowest RSD and toluene (RSD = 528%) the highest. These relatively large RSDs are attributed to a limited number of samples with high exposure to BTEXS that are not categorized as crude oil/fuel. These samples did not appear to have the characterized crude oil/fuel exposure pattern, but may be the result of misclassification or exposure to a VOC source not associated with crude oil/fuel or smoke. For example, one sample (SEQN# 51422) classified as nonsmoker had only high toluene (24.1 ng/mL or ppb) with all other monoaromatic compounds below 0.210 ng/mL. The blood 2,5-dimethylfuran level for this sample was 0.015 ng/mL, which is above the previously demonstrated smoking biomarker cutpoint of 0.014 ng/mL, but the blood toluene level is more than an order of magnitude higher than any smoker. The ANN classification of this sample as other/nonsmoker demonstrates the strength of using a multianalyte signature in making a categorization determination rather than considering just one biomarker level. Removing this one sample decreased the RSD for toluene from 528% to 235%. The variability for the crude oil/fuel group ranged from 42% to 233% and was lowest for benzene (RSD = 42%) and highest for styrene (RSD = 233%). These results show that individuals categorized as crude oil/fuel can have differing exposure levels.

5. Conclusions

Based on this work and previous work in our laboratory, petroleum and petroleum-based products can produce a BTEX signature distinguishable from other sources such as tobacco smoke even though these sources are composed of many of the same compounds. When combined with other concomitant VOCs such as styrene and 2,5-dimethylfuran, BTEX signatures associated with crude oil/fuel can be reliably distinguished from signatures related to tobacco smoke or those not consistent with to either of these sources. The use of multiple biomarkers minimizes the effect of confounding from other sources that can otherwise occur with relying on a single biomarker cutpoint level. Unique to this work is the

use of surrogate signatures derived by adjusting the source VOC composition with respective blood/air partition constants and normalization of the VOC levels to a reference VOC with high detection frequency. Although best practice would be to use actual exposed individuals, use of a surrogate for petroleum exposure was necessary because sufficient exposure data were not available. A limitation of using surrogate signatures is that there is no magnitude information, only relative level. Without magnitude, surrogate signatures cannot be adjusted to take into account baseline levels. Absence of this information can cause low-level exposure to be classified as other/nonsmoker as that signature more closely reflects the baseline signature associated with no exposure.

The use of ANN for pattern recognition provided the means to reliably identify individuals with signatures consistent with those used to train the ANN, which is particularly useful when working with large study populations such as NHANES. Furthermore, in performing categorization with ANN, occurrence between different populations could be assessed objectively because fixed model calculations are universally applied. Using the ANN, 7 individuals out of 1998 (0.35%) from NHANES 2007–08 and 12 individuals out of 2906 (0.41%) had blood signatures consistent with the crude oil/fuel surrogate signature. Visual inspection of these crude oil/fuel signatures confirms the identification made by the ANN model. Based on these analyses, blood VOC levels in the United States indicated that substantial exposures to crude oil and fuel are infrequent.

Acknowledgments

The authors wish to thank Olivia Harris for detailed comments and Nathan Geldner for helpful edits involving the statistical aspect of this manuscript. Lydia Thornburg, Eduardo Sanchez, Jessica Rafson, and John Ruhl III were funded by the Research Participation Program at the Centers for Disease Control and Prevention, an interagency agreement with the U.S. Department of Energy administered by the Oak Ridge Institute for Science and Education.

Reference List

Anderson JC, Babb AL, Hlastala MP. Modeling soluble gas exchange in the airways and alveoli. Ann Biomed Eng. 2003; 31:1402–1422. DOI: 10.1114/1.1630600 [PubMed: 14758930]

Ashley DL, Bonin MA, Hamar B, McGeehin M. Using the blood concentration of 2,5-dimethylfuran as a marker for smoking. Int Arch Occ Env Hea. 1996; 68:183–187.

Bialous SA, Yach D. Whose standard is it, anyway? How the tobacco industry determines the International Organization for Standardization (ISO) standards for tobacco and tobacco products. Tob Control. 2001; 10:96–104. DOI: 10.1136/tc.10.2.96 [PubMed: 11387528]

Blount BC, Kobelski RJ, McElprang DO, Ashley DL, Morrow JC, Chambers DM, et al. Quantification of 31 volatile organic compounds in whole blood using solid-phase microextraction and gas chromatography-mass spectrometry. J Chromatogr B Analyt Technol Biomed Life Sci. 2006; 832:292–301. DOI: 10.1016/j.jchromb.2006.01.019

Centers for Disease Control and Prevention (CDC).National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: US Department of Health and Human Services, Centers for Disease Control and Prevention; 2007–08. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2007 [Last accessed 7/6/2017]

Centers for Disease Control and Prevention (CDC).National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: US Department of Health and Human Services, Centers for Disease Control and Prevention; 2013–14. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2013 [Last accessed 7/6/2017]

Chambers DM, Blount BC, McElprang DO, Waterhouse MG, Morrow JC. Picogram measurement of volatile n-alkanes (n-hexane through n-dodecane) in blood using solid-phase microextraction to assess nonoccupational petroleum-based fuel exposure. Anal Chem. 2008; 80:4666–4674. DOI: 10.1021/ac800065d [PubMed: 18481873]

Chambers DM, McElprang DO, Waterhouse MG, Blount BC. An improved approach for accurate quantitation of benzene, toluene, ethylbenzene, xylene, and styrene in blood. Anal Chem. 2006; 78:5375–5383. DOI: 10.1021/ac060341g [PubMed: 16878872]

Chambers DM, Ocariz JM, McGuirk MF, Bount BC. Impact of cigarette smoking on Volatile Organic Compound (VOC) blood levels in the U.S. Population: NHANES 2003–2004. Environ Int. 2011; 37:1321–1328. DOI: 10.1016/j.envint.2011.05.016 [PubMed: 21703688]

Chawla NV, Bowyer KW, Hall LO, Kegelemeyer WP. SMOTE: Synthetic Minority Over-Sampling Technique. J Artif intell Res. 2002; 16:321–357. DOI: 10.1613/jair.953

Environment Canada Environmental Technology Centre. [Last accessed 7/6/2017] Oil Properties Database. 2001. http://www.etc-cte.ec.gc.ca/databases/oilproperties

Gordon SM. Identification of exposure markers in smokers' breath. J Chromatogr. 1990; 511:291–302. [PubMed: 2211914]

Gotwalt, C. JMP© 9 Neural Platform Numerics. Cary, NC: SAS Institute Inc; 2011. https://www.jmp.com/content/dam/jmp/documents/en/white-papers/wp-jmp9-neural-104886.pdf [Last accessed 6/7/2017]

Grandjean P, Landrigan PJ. Neurobehavioural effects of developmental toxicity. Lancet Neurol. 2014; 13:330–338. DOI: 10.1016/S1474-4422(13)70278-3 [PubMed: 24556010]

Health Canada. Determination of "Tar", Nicotine and Carbon Monoxide in Mainstream Tobacco Smoke. Health Canada Offical Publication; 1999. T-155

Jakubowski M, Czerczak S. Calculating the retention of volatile organic compounds in the lung on the basis of their physicochemical properties. Environ Toxicol Pharmacol. 2009; 28:311–315. DOI: 10.1016/j.etap.2009.05.011 [PubMed: 21784021]

Jamal A, Homa DM, O'Connor E, Babb SD, Caraballo RS, Singh T, Hu SS, King BA. Current Cigarette Smoking Among Adults - United States, 2005–2014. Centers for Disease Control and Prevention. 2015; 64(44):1233–1240.

Kim KH, Jahan SA, Kabir E. A review of breath analysis for diagnosis of human health. TrAC Trends in Anal Chem. 2012; 33:1–8. DOI: 10.1016/j.trac.2011.09.013

Korte JE, Hertz-Picciotto I, Schulz MR, Ball LM, Duell EJ. The contribution of benzene to smoking-induced leukemia. Environ Health Persp. 2000; 108:333–339.

Kramer C, Mochalski P, Unterkofler K, Agapiou A, Ruzsanyi V, Liedl KR. Prediction of blood:air and fat:air partition coefficients of volatile organic compounds for the interpretation of data in breath gas analysis. J Breath Res. 2016; 10:017103.doi: 10.1088/1752-7155/10/1/017103 [PubMed: 26815030]

Liu H, Man H, Tschantz M, Wu Y, He K, Hao J. VOC from Vehicular Evaporation Emissions: Status and Control Strategy. Environ Sci Technol. 2015; 49:14424–14431. DOI: 10.1021/acs.est.5b04064 [PubMed: 26599318]

Mallach G, St-Jean M, MacNeill M, Aubin D, Wallace L, Shin T, et al. Exhaust ventilation in attached garages improves residential indoor air quality. Indoor Air. 2017; 27:487–499. DOI: 10.1111/ina.12321 [PubMed: 27444389]

Maltoni C, Conti B, Cotti G, Belpoggi F. Experimental Studies on Benzene Carcinogenicity at the Bologna Institute of Oncology - Current Results and Ongoing Research. Am J Ind Med. 1985; 7:415–446. [PubMed: 4003403]

Meulenberg CJ, Vijverberg HP. Empirical relations predicting human and rat tissue:air partition coefficients of volatile organic compounds. Toxicol Appl Pharmacol. 2000; 165:206–216. DOI: 10.1006/taap.2000.8929 [PubMed: 10873711]

Mogel I, Baumann S, Bohme A, Kohajda T, von Bergen M, Simon JC, et al. The aromatic volatile organic compounds toluene, benzene and styrene induce COX-2 and prostaglandins in human lung epithelial cells via oxidative stress and p38 MAPK activation. Toxicology. 2011; 289:28–37. DOI: 10.1016/j.tox.2011.07.006 [PubMed: 21801798]

Pazo DY, Moliere F, Sampson MM, Reese CM, Agnew-Heard KA, Walters MJ, et al. Mainstream Smoke Levels of Volatile Organic Compounds in 50 U.S. Domestic Cigarette Brands Smoked With the ISO and Canadian Intense Protocols. Nicotine Tob Res. 2016; 18:1886–94. DOI: 10.1093/ntr/ntw118 [PubMed: 27113015]

Potter, TL., Simmons, KE. Composition of petroleum mixtures. Amherst, Mass: Amherst Scientific Publishers; 1998. Total Petroleum Hydrocarbon Criteria Working Group.

Sampson MM, Chambers DM, Pazo DY, Moliere F, Blount BC, Watson CH. Simultaneous Analysis of 22 Volatile Organic Compounds in Cigarette Smoke Using Gas Sampling Bags for High-Throughput Solid-Phase Microextraction. Anal Chem. 2014; 86:7088–7095. DOI: 10.1021/ac5015518 [PubMed: 24933649]

Symanski E, Stock TH, Tee PG, Chan W. Demographic, residential, and behavioral determinants of elevated exposures to benzene, toluene, ethylbenzene, and xylenes among the U.S. population: results from 1999–2000 NHANES. J Toxicol Environ Health A. 2009; 72:915–924. DOI: 10.1080/15287390902959706 [PubMed: 19557620]

United States Environmental Protection Agency. Volatile Organic Compounds Emissions. United States Environmental Protection Agency; 2014.

Wagner PD, Saltzman HA, West JB. Measurement of continuous distributions of ventilation-perfusion ratios: theory. J Appl Physiol. 1974; 36:588–599. [PubMed: 4826323]

Wang, Z., Hollebone, BP., Fingas, M., Fieldhouse, B., Sigouin, L., Landriault, M., et al. US EPA Report EPA/600/R-03/072. Jul. 2003 Characteristics of spilled oils, fuels, and petroleum products: 1. Composition and Properties of Selected Oils.

Xu XH, Freeman NC, Dailey AB, Ilacqua VA, Kearney GD, Talbott EO. Association between Exposure to Alkylbenzenes and Cardiovascular Disease among National Health and Nutrition Examination Survey (NHANES) Participants. Int J Occup Env Heal. 2009; 15:385–391. DOI: 10.1179/oeh.2009.15.4.385

Yamada H. Contribution of evaporative emissions from gasoline vehicles toward total VOC emissions in Japan. Sci Total Environ. 2013; 449:143–149. DOI: 10.1016/j.scitotenv.2013.01.045 [PubMed: 23422493]

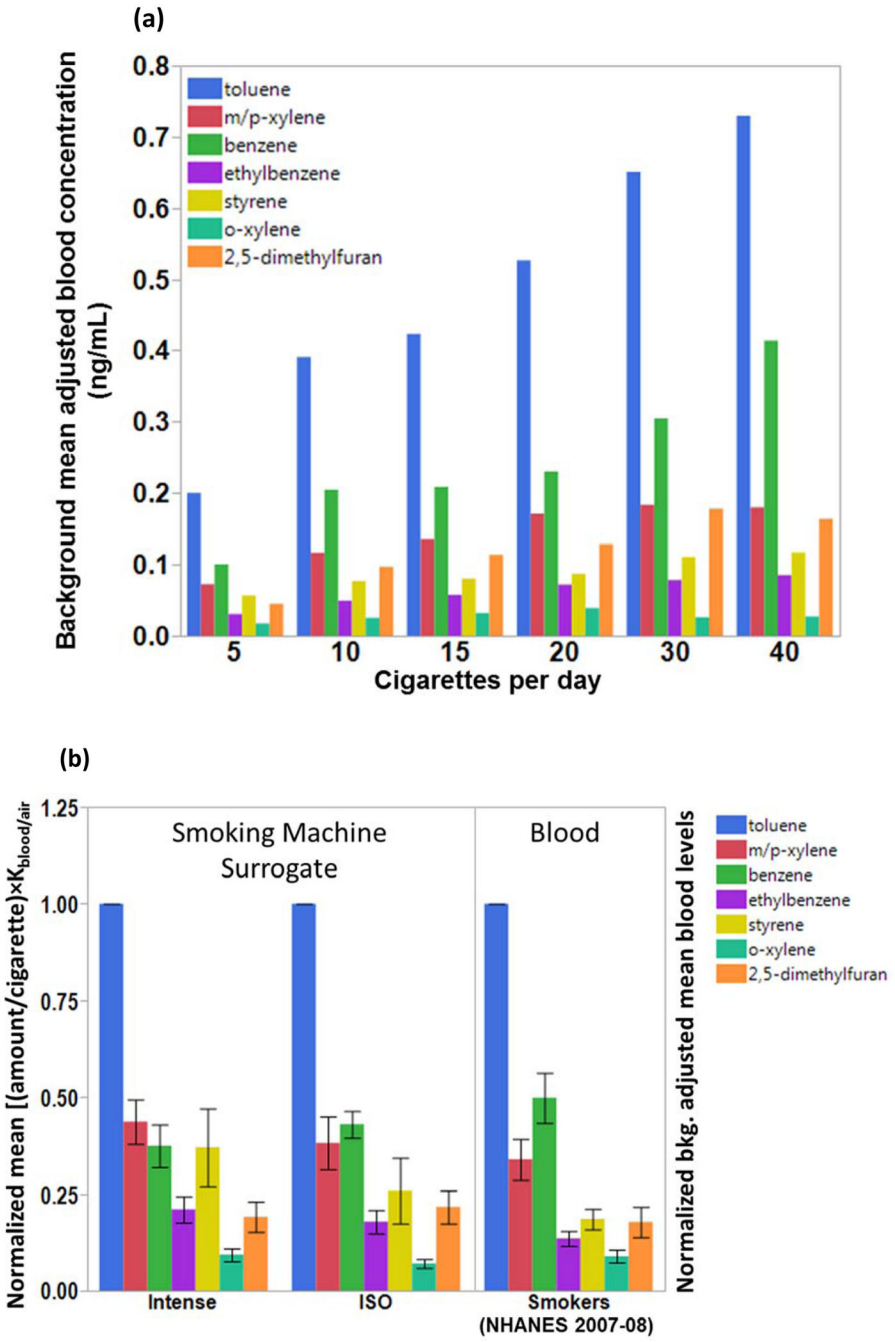

Figure 1.

Blood BTEXS signature among smokers remains consistent despite demographic and smoking technique differences, although absolute levels vary substantially. (a) Comparison of baseline level adjusted mean blood concentrations of BTEXS and 2,5-dimethylfuran among cigarette smokers who report smoking 5 (N=47), 10 (N=71), 15 (N=42), 20 (N=91), 30 (N=25), and 40 (N=14) cigarettes per day from the 2007–08 NHANES. (b) Comparison of normalized (relative to toluene) and adjusted machine generated cigarette smoke signatures using Canadian Intense and ISO protocols from 50 U.S. brand varieties with a

composite of blood levels among all NHANES 2007–08 cigarette smokers. Cigarette smoke signatures comprise BTEXS and 2,5-dimethylfuran levels adjusted by the corresponding blood/air partition constant ($K_{blood/air}$). Error bars are 1 standard deviation from the mean.

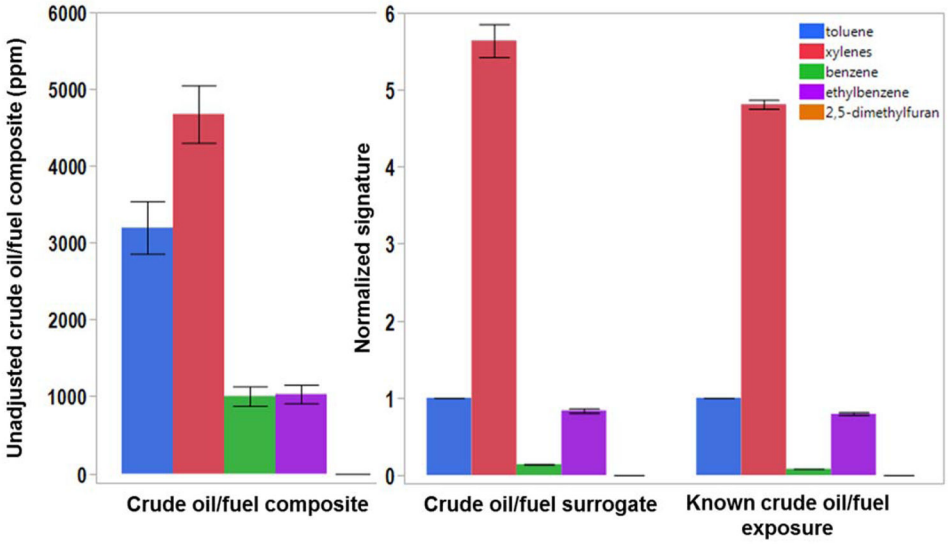

Figure 2.

Comparison of composite petroleum signature (25 crude oils and 10 fuels) and surrogate blood level, with the blood signature from a known fuel exposure demonstrating similarity between surrogate and known exposure. Error bars are 1 standard deviation from the mean.

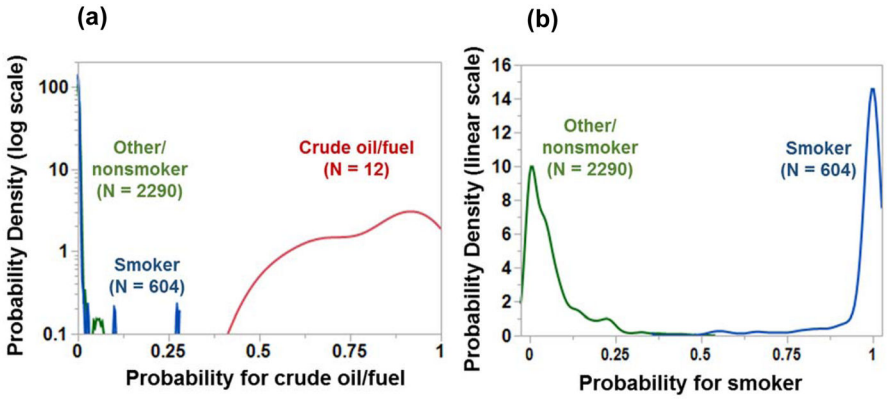

Figure 3.

Comparison of probability density from artificial neural network showing how well categorization is defined between the different categories for (a) crude oil/fuel exposure vs. other/nonsmoker and smoker and for (b) tobacco other/nonsmoker vs. smoker for the NHANES 2013–14 study data.

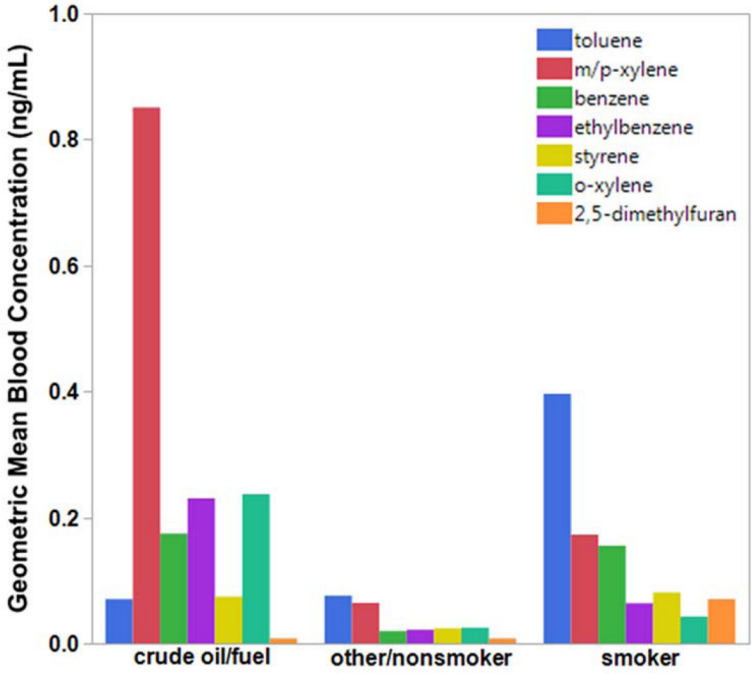

Figure 4.

Comparison of BTEXS/2,5-dimethylfuran blood signature composites categorized by artificial neural network for crude oil/fuel, nonsmoker, and smoker groups from NHANES 2007–08.