MATHEMATICAL AND STATISTICAL APPROACHES FOR INTERPRETING BIOMARKER COMPOUNDS IN EXHALED HUMAN BREATH

JOACHIM D. PLEIL AND JON R. SOBUS

National Exposure Research Laboratory, Office of Research and Development
U.S. Environmental Protection Agency, Research Triangle Park, NC 27711 USA

1. Introduction

The various instrumental techniques, human studies, and diagnostic tests that produce data from samples of exhaled breath have one thing in common: they all need to be put into a context wherein a posed question can actually be answered. Exhaled breath contains numerous compounds; just the volatile organic fraction alone has been estimated to represent in excess of 500 different chemical species. In addition, the aerosol fraction contains proteins, signaling molecules, dissolved inorganic compounds, and even bacteria and viruses adding to the complexity of the total sample.

No single technique can detect everything in breath, in fact, even the most broadly designed breath measurements result in suites of compounds restricted by the methods used. For example, reactive oxygen species may be observed using real-time sensors or real-time mass spectrometry (MS), but not by gas chromatography – MS (GC-MS), whereas GC-MS can discriminate among a variety of hydrocarbons, alcohols, and ketones that may overlap completely in a real-time MS instrument without benefit of chromatographic separation. Furthermore, the fraction of the breath (gas-phase or aerosol phase) also determines what measurements can be made; for example proteins and signaling molecules could be detected in exhaled breath condensate via enzyme-linked immuno-sorbent assay (ELISA), nuclear magnetic resonance (NMR), or liquid chromatography (LC) MS, but not with any gas-phase instruments such as those based on optical spectroscopy or gas chromatography.

The first issue that the data analyst faces is that all data he or she sees, regardless how complex or detailed, is not comprehensive but always stratified (restricted) by the chemistry, instrumentation, and thermodynamics

of the choices made for sampling and analysis. The second issue is that a suite of compounds measured in any given sample does not represent an independent set of variables; sub-groups of biomarker compounds often have appreciable covariance reflecting similar metabolic pathways or exogenous sources. The third issue relates to variability; any given sample is unique and it can never be taken for granted that the constituents of the breath are the same between people nor within one person over time. Finally, any individual compound in breath can have a wide range of concentrations that are all considered "normal" or "unremarkable" in the apparently healthy general population. This has implications for assessing health or exposure status based on just a few data points; under such constraints, the analyst can only interpret a measurement as a statistical probability that the concentration is probative.

2. Data interpretation

In this chapter, we describe a series of mathematical and statistical approaches geared specifically to the interpretation of volatile organics in exhaled breath that can be implemented to address the four issues outlined above. Although all methods interact, we have assigned them to five categories for the purposes of this discussion as follows:

(1) Data visualization and summary statistics
(2) Variable independence and clustering
(3) Population statistics and variance components
(4) Stochastic models and meta-data
(5) Dynamic models and longitudinal data

Each of these approaches can provide distinct information about a breath data set, generally increasing in detail in the order listed. We note that the broader interpretations gleaned from categories 1, 2, and 3 feed the modeling processes in 4 and 5. We further note that not all of these procedures need to be performed; often it is sufficient to answer a question with a simple analysis based on a graph or a summary table. In the following, we describe these categories and show how they can progressively tell a story about acquired data. We assume that multiple compounds have been measured across a number of people, subjects may be grouped as cases and controls, host factor meta-data have been collected, and measurements may have been repeated, either with or without intervention or treatment.

2.1. *Data visualization and summary statistics*

The first step in any analysis of newly acquired data is to get a "feel" for how the experiment turned out. Generally, we make graphs and calculate averages, standard deviations, data ranges, etc. We have found that it is very helpful to get all of the data into some form of visual representation, either as bar graphs, cluster plots, or heat maps, depending on complexity of the data structure. For example, in a recent publication, we described measurements of a series of polar volatile organic compounds (PVOCs) in exhaled breath condensate (EBC) made during an intervention study of diesel exhaust exposure.[1] Human subjects, (3 males, 6 females) were each exposed for 2 hrs to a dilute diesel exhaust atmosphere and to a purified air atmosphere on separate occasions.[2] A variety of samples, including EBC, were collected immediately before and after the exposures, and again 24 hrs later; EBC means data for eight of the most prevalent compounds were plotted by nine subjects grouped by gender (Figure 1). We noted a slight gender bias for some compounds, but otherwise the results were unremarkable and we went on to assess the data from a statistical perspective. In a subsequent publication, however, we developed methodology for data

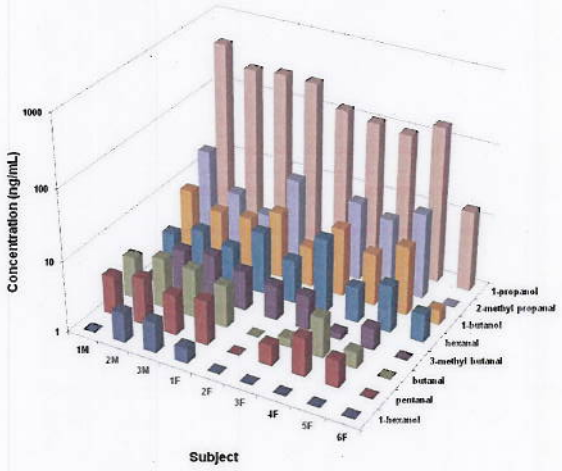

Figure 1. Bar graph visualization of balanced exhaled breath data from an intervention study of diesel exhaust exposure. Summary breath data for nine subjects and eight compounds.

visualization using heat map style graphics.[3] Briefly, heat maps are visual representations of quantitative data on two axes; the x-axis reflects individual samples and the y-axis consists of groups of measured parameters. The field between the axes is comprised of an array contiguous boxes color coded to reflect quantitation. The term "heat map" is derived from the convention that the higher levels tend towards red, the lower levels tend towards blue.

We revisited the data set from Hubbard et al.[1] and found that heat map visualization was capable of showing more pattern detail. This technique had to be restricted to seven subjects with complete data to avoid blank spots on the heat map as we lost a few samples to follow-up. Figure 2 shows this alternate approach; note that we now have access to results from individual subjects' samples, for all 16 measured PVOCs, and the flexibility for grouping samples by gender and longitudinal parameters. Here, the gender effect becomes obvious; males were expressing much higher levels of many of the measured PVOCs, especially 2-methyl propanal, 1-heptanol, butanal, pentanal, 1-hexanol, and 3-methyl-3-pentanol. Only hexanal and heptanal reverse this trend. We further see that there is no apparent treatment effect or longitudinal time effect, that is, there are no obvious pattern differences

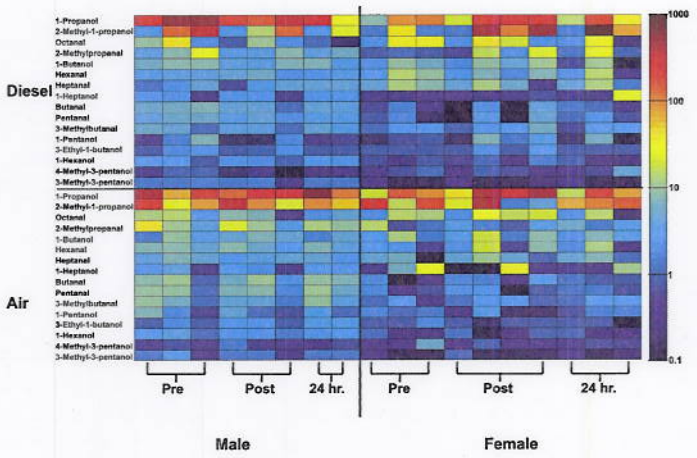

Figure 2. Heat map visualization of balanced exhaled breath data from an intervention study of diesel exhaust exposure. Individual samples grouped by longitudinal time frame (pre-exposure, post- exposure and 24 hr post-exposure) as well as by gender.

between diesel exposures and air exposures, nor pre- and post-exposure data. These simple visualization approaches, coupled with summary statistics, can provide hints as to how more detailed mathematical approaches could be implemented to quantify these observations. Also, we note that the visualization methods for complex data make the subsequent statistical calculations more accessible to the readership.

2.2. *Variable independence and clustering*

One of the most vexing problems encountered in data interpretation procedures is the determination of actual independence of what we generally denote as "independent variables". Subsequent statistical evaluations rely on the notion that measurements that are used in models are not overly correlated. For example, consider the height and weight parameters of human subjects as host factor data in a complex breath experiment. We know that taller people tend to be heavier, and so there is significant correlation expected. If one were to treat height and weight as independent and place them both into a model for predicting a health outcome (along with a series of breath biomarkers), one can get completely different interpretations as to their respective importance depending upon which was entered first. This why one often sees body-mass index (BMI) used as a composite parameter instead. Now consider unexpected correlations among variables, for example, among compounds measured in the exhaled breath of subjects in an intervention or case-control study. If certain compounds are repeatedly behaving the same way, then their inclusion in even simple multivariate regression models will result in mathematical instability.

We cannot know *a priori* if different analytes are tracking the same outcome, however, we can perform various correlation tests among presumed independent variables to assess the degree of independence. The most fundamental method is the correlation matrix wherein the regression between pairs of variables (V_i, V_k) is calculated as the Pearson correlation coefficient "r" which has possible values ranging from -1 to 1. Positive r-values indicate that a larger V_i is associated with a larger V_k, whereas a negative r indicates that a larger V_i is associated with a smaller V_k. The closer the r-value is to 0, the more independent the two variables. For example, in a study measuring height (H), weight (W), and percent body fat (BF), $r = 0.486$ for W vs H, $r = 0.074$ for H vs BF, and $r = 0.613$ for W vs BF.[4] Based on the results of this simple correlation matrix, one would probably

not include both parameters H and W, nor both W and BF, but would keep H and BF as host factors in a resulting model.

The correlation matrix approach is considered a semi-quantitative measure for overall data independence interpretation because it treats variables two at a time; one cannot discern directly how multiple variables, or combinations of variables, correlate. A more sophisticated approach is available wherein Eigen-vector projections are calculated for all variables in n-dimensional space. Using statistical software such as Proc VarClus from SAS Inc. (Cary, NC, USA), it is possible to develop a "dendrite" diagram that can be used to create clusters of variables that have certain levels of auto-correlation. This is similar to principle components analysis (PCA), but rather than grouping samples, this approach groups variables. Variable clustering serves two purposes, it improves model stability by identifying/removing collinearity, and reduces the total number of variables for a more parsimonious model.[5]

As an example, consider the dendrite diagram (dendrogram) in Figure 3 where we show a generic example of both host factor and measurement data

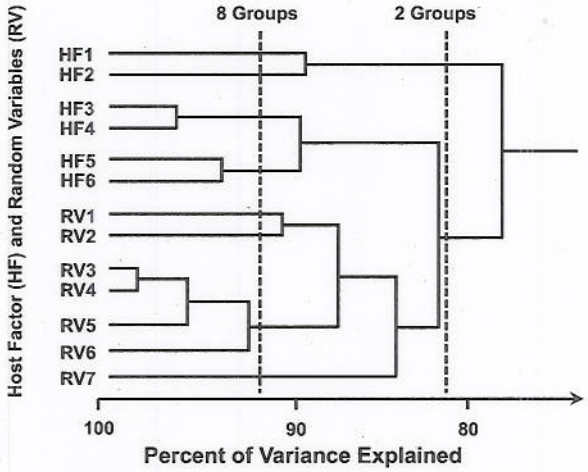

Figure 3. Example dendrogram of variable cluster analysis. Starting with a total of $m = 13$ independent variables ($p = 6$ host factors and $q = 7$ environmental variables), forming 8 clusters (HF1, HF2, HF3+HF4, HF5+HF6, RV1, RV2, RV3+RV4+RV5+RV6, and RV7) results in explained variance of $\sim$ 93%. Collapsing the clusters further to 2 explains $\sim$ 83% of the variance, but now is difficult to interpret.

clustering.[5] In this case, we use a hypothetical data structure with 6 host factors (gender, BMI, age, etc.) and 7 environmental/biomarker variables (e.g. concentration measurements in blood or urine). This methodology has been applied to complex environmental dioxin congener data[6] and to studies of jet fuel exposure in the US Air Force.[7]

By collapsing the more highly correlated variables into clusters, we do not lose much explanatory power, but improve the interpretive power of subsequent models by increasing the ratio (n/m) between number of samples "n" and the number of independent variables "m". We have observed that $n/m \geqslant 10$ is a good rule of thumb for assessing the importance of individual variables.[6] We caution that this procedure requires a certain amount of judgment on the part of the investigator. The first issue is choosing the variance level vs. number of clusters. This is very dependent on the eventual n/m parameter. The second issue is how to combine information from variables collapsed into clusters. There are a number of choices that depend on the nature of the variables; in the above example, we use a simple sum, but other strategies can be employed as well. For example, given highly correlated variables such as HF3 and HF4 (especially if they are not continuous), one can just discard one or the other. For measured concentrations, it is sometimes better to normalize each measurement to a total to avoid having a particular variable overwhelm the sum.

2.3. *Population statistics and variance components*

Traditional mathematical analyses of breath biomarkers, both for clinical and environmental research applications, utilize measurement statistics (e.g., mean and median) to evaluate differences between groups. For example, it is common to evaluate measurements statistics to compare cases vs. controls, exposed vs. unexposed subjects, or males vs. females (as shown in our earlier example of a diesel intervention study). Furthermore, it is common to employ stratified data analyses – that is, analyses using multiple levels of organization – to reduce the impacts of confounding variables. Consider the evaluation of smoking effects on breath biomarker levels; one could perform a single evaluation using all subjects (smokers vs. non-smokers), two evaluations after stratifying by sex (male smokers vs. male non-smokers; female smokers vs. female non-smokers), four evaluations after stratifying by sex and disease status (male smoker [case] vs. male smoker [control]; male non-smoker [case] vs. male non-smoker [control]; female smoker [case] vs. female smoker [control]; female non-smoker [case] vs. female non-smoker

[control]), and so on. These stratified analyses can narrow down the potential origins of an observed effect, but require larger sample numbers with each level of stratification. Thus, an investigator must balance the desired outcome of a given analysis with costs required to achieve sufficient statistical power and sensitivity.[6]

Given adequate sample numbers for stratified group analyses, it is prudent to first investigate the underlying distribution(s) of the biomarker data in question. Our earlier examples for data analysis (i.e., data visualization and variable clustering) can be considered qualitative or semi-quantitative, and generally have no *a priori* conditions for data structure. Hypothesis testing on the other hand, is entirely quantitative, and in many cases relies on distributions of measurement data being approximately Gaussian, or "normal". Simple diagnostic procedures exist in most software packages to evaluate data distributions. Often times, simple histograms, normal probability plots, or quantile-quantile plots can be visually inspected to evaluate normality assumptions. More advanced statistical packages offer statistical tests (e.g., Shapiro-Wilk and Kolmogorov-Smirnov) to confirm results from visual inspection.

Measurements of analytes in biological media are often "log-normally" distributed; that is, the distribution of the logged values is approximately normal. Log-normal data can be identified by a right-skewed distribution of the original (non-transformed) data, reflecting few values at exceedingly high levels, and many values near, but not below, a lower threshold (generally zero). Additional signs of log-normally distributed data include (1) an observed arithmetic mean value that is larger than the median (reflecting the differential effects of extremely high values on these statistical parameters), and (2) an observed confidence interval that includes negative values (it is impossible to have negative amounts of a biomarker). When one encounters these signs, it is best to evaluate data transformation approaches before pursuing quantitative hypothesis testing.

"Parametric" testing procedures are generally used to evaluate normally distributed data and log-normally distributed data that have been log-transformed. Commonly used parametric tests include the Student's t-test, the paired t-test, and analysis of variance (ANOVA). A Student's t-test evaluates differences in biomarker measurements across two groups (e.g. men vs. women); a paired t-test evaluates differences in biomarker measurements between paired samples from individuals (e.g. pre-intervention vs. post-intervention); and ANOVA compares biomarker measurements across multiple groups (e.g. children vs. adolescents vs. adults). In the event that

biomarker data are not Gaussian, or cannot be mathematically transformed to approximate normality, "non-parametric" methods exist for hypothesis testing. For example, the Mann-Whitney test, Wilcoxon test, and Kruskal-Wallis test are common substitutes for the Student's t-test, paired t-test, and ANOVA, respectively. These tests generally approximate the results of the parametric tests given a large enough sample size. However, the exact data requirements (e.g., sample number, sample independence, random selection) for any given test should always be considered prior to hypothesis testing.

The statistical tests discussed thus far generally utilize a single observation for each subject. (While the paired t-test uses two observations per subject, the difference between the two values [i.e., a single value] is used for hypothesis testing). Therefore, different approaches to data analysis are required when multiple measurements exist for each subject. At the most basic level, it is of interest to evaluate the spread of the data, or the "variance", between and within subjects. To do this, the total variance across all measurements is first partitioned into that which is observed across repeated measurements of individual subjects, known as "within-person (intra-individual) variance", and that which is observed across average levels of all subjects known as "between-person (inter-individual) variance". These between- and within-person variance components can be calculated using a number of techniques.[8] The simplest of these methods is analysis of variance (ANOVA) which is suitable for balanced data sets where the same number of measurements exists for each subject. More complex methods, such as restricted maximum likelihood estimation (REML), may be required when working with unbalanced data sets; these methods are typically available only in more advanced statistical software packages.

Once the variance components are established, they become instrumental for identifying the origins of exposure or disease, and in turn, the best opportunities for mitigation or intervention. Large within-person variance and small between-person variance indicates little difference between subjects on average, but large differences over time for each subject. This result may point to a temporal event, random or otherwise, that affects each subject equally. Alternatively, small within-person variance and large between-person variance indicates little change over time, but marked differences between subjects. This result may point to a host-specific parameter (e.g., genotype, fitness-level) that dictates biomarker response. In the first example, a general intervention strategy, equally applicable to all subjects, might be suitable to reduce an exposure or limit a biological response.

In the second example, a targeted strategy would likely be necessary to first identify the cause of elevated biomarker levels for certain individuals, and then for intervention.

2.4. *Stochastic models and meta-data*

After identifying group-based differences in biomarkers levels, and within- and between-subject variance components (for repeated-measures studies only), the next step is to develop statistical models using study meta-data. Statistical models serve two functions for breath research; (1) they allow investigators to simultaneously identify multiple significant predictors of breath biomarker levels; and (2) they provide a platform for predicting breath biomarker levels in other studies where only meta-data exist.

An important decision for model development is identifying a dependent variable; this is not as easy as it sounds. Often, the dependent variable is obvious by design and represents the "outcome" for a subject; it can be binomial, that is, cancer/not cancer, infected/not infected, or it can be a continuous health outcome variable such as cholesterol level, total urinary protein, pulmonary function (e.g. forced expiratory volume in 1 sec (FEV1), forced vital capacity (FVC)), or DNA damage (e.g. sister chromatid exchange (SCE), strand breaks), among others.

However, there are many occasions when the choice of the dependent variable is not obvious, especially in environmental or cross-sectional public health studies for which the analyst was not consulted in sampling design. For example, suppose one has measured a suite of exogenous chemicals and metabolites in exhaled breath, and has acquired meta-data concerning recent activity, job type, height, weight, gender, ethnicity, etc., but no health effects information was collected or observed. What could the dependent variable be for interpretation purposes? In such cases, the first question to be addressed is: What do we want to know? Generally, we would like to explore the linkages between suspected exposure sources and the resulting internal dose in human subjects. Suppose that in the totality of all measurements, we observe benzene, toluene, ethyl-benzene and xylenes (BTEX), plus many other organic compounds in breath. If one of the suspected sources for all exposures is automobile exhaust, we could sum the BTEX numbers into one composite parameter and use this to represent the dependent variable for overall fuels exposure. This is a bit of a bootstrapping approach, especially if we keep the individual BTEX compounds as

independent continuous variables, but often this is the only way to build a stochastic model in the absence of a designed dependent variable.

Given that we have a number of "independent variable" measurements and related meta-data and some consensus dependent (outcome) variable, then the next step is to determine which independent variables and data actually tell a story relating to the changes in the identified outcome variable. This requires some form of a modeling approach, and if the data include both host factors and continuous variables, the best approach is a generalized linear model, or a "mixed effects" model.[8,9]

A general form of such a model is:

$$Y_{hij} = \left\{\beta_1 X_{1_{hij}} + \beta_2 X_{2_{hij}} + \cdots + \beta_p X_{p_{hij}}\right\} + \alpha_h + b_{hi} + \epsilon_{hij}$$

where:

- Y_{hij} is the value of some relevant biological parameter for the j^{th} observation of the i^{th} subject in the h^{th} group;
- $X_{1_{hij}}, X_{2_{hij}}, \ldots, X_{p_{hij}}$ are the values of the fixed effect variables such as environmental chemical concentrations (in air, water, food, dust, etc.), and host factors such as age, health state, gender, genetic polymorphisms, ethnicity, etc.;
- p is the total number of fixed effects (note: the host factors may be fixed for all j within a given i);
- $\beta_1, \beta_2, \ldots, \beta_p$ are the corresponding modeled coefficients for the fixed effects and host factors;
- α_h is the random effect for the h^{th} group;
- b_{hi} is the random effect for the i^{th} subject from the h^{th} group;
- ϵ_{hij} is the residual (unexplained) error for the j^{th} observation of the i^{th} subject from the h^{th} group.

Software applications for this style of approach are commercially available (e.g. proc MIXED, SAS). Upon calculation, the coefficients and their p-values and can be interpreted to determine the effect of including the particular fixed effect or random effect variable in the final model for explaining the variance in the biological parameters' values. This can be done with iterative steps of forward addition or reverse elimination with the eventual objective being a parsimonious model without appreciable loss of modeling power. Once the final model is established, we can observe which exposure parameters and fixed effects are more likely to cause perturbations to the systems biology.[5]

In an earlier example, based on published results from a diesel exhaust intervention study, we demonstrated how visualization tools can be used to identify likely predictors of breath biomarker levels. Specifically, we used a 3D bar graph and a heat map to show the likely effect of gender, and the unlikely effects of the exposure intervention and sample time, on PVOC levels in EBC. Following from this qualitative work, we now demonstrate a statistical assessment of this data using mixed models. We note that these methodologies and results have been published alongside the aforementioned graphical work.[1]

PVOC levels in breath were treated individually as model dependent variables, whereas model independent variables included fixed effects for the time of sample collection (pre-intervention vs. post-intervention vs. 24-hr post-intervention), the type of intervention (diesel exhaust vs. purified air), and gender (male vs. female), as well as random effects for subject and residual error. Results of the models are shown in Table 1, and confirm a significant gender effect ($p < 0.1$) for four out of nine modeled PVOCs, with men having higher levels than women in all cases. While not shown in Table 1, results also confirm no significant effects ($p > 0.1$) of collection time or exposure intervention on these PVOCs. These quantitative results corroborate those of the earlier qualitative analyses. Furthermore, these results include variance components estimates (listed under "Random effects" in Table 1) that can be used to explore subject susceptibilities and possible intervention strategies.[10,11]

2.5. *Dynamic models and longitudinal data*

Dynamic models are used to interpret time dependence. In some study designs, the breath data structure investigates certain applied or observed external conditions. In diagnostic medicine, this could be a pre/post drug treatment study, or a time-dependent study to monitor post-operative recovery status. In environmental studies, this could be a sample time-series to determine the rates of uptake and elimination from different profiles of deliberate or incidental exposures to chemicals. Regardless of the exact design, the primary focus is temporally resolved data. Such studies are invaluable for deducing the time constants of absorption, distribution, metabolism and elimination of exogenous compounds like pharmaceuticals and environmental pollutants, and for assessing different metabolic pathways. The underlying assumptions of such a study are that kinetic and physiological parameters measured under controlled conditions will be con-

Table 1. Mixed models results for PVOCs in EBC from subjects enrolled in a diesel exhaust intervention study.

PVOC	Fixed effects						Random effects					
	Intercept			Gender			$\hat{\sigma}_b^2$			$\hat{\sigma}_w^2$		
	Est.	(*SE*)	*p-value*	*Est.*[1]	(*SE*)	*p-value*	*Est.*	(*SE*)	*p-value*	*Est.*	(*SE*)	*p-value*
1-hexanol	−0.592	(0.281)	0.07	1.27	(0.483)	0.01	0.438	(0.251)	0.04	0.138	(0.033)	< 0.0001
butanal	0.135	(0.355)	NS	1.31	(0.602)	0.04	0.583	(0.406)	0.08	0.710	(0.170)	< 0.0001
pentanal	0.151	(0.347)	NS	1.27	(0.588)	0.04	0.545	(0.382)	0.08	0.734	(0.175)	< 0.0001
2-methylpropanal	1.33	(0.273)	0.002	0.787	(0.453)	0.09	0.206	(0.240)	NS	1.04	(0.248)	< 0.0001
1-propanol	4.86	(0.304)	< 0.0001			NS	0.733	(0.411)	0.04	0.450	(0.107)	< 0.0001
1-butanol	1.63	(0.191)	< 0.0001			NS	0.274	(0.161)	0.05	0.239	(0.058)	< 0.0001
hexanal	1.32	(0.220)	0.0003			NS	0.325	(0.209)	0.06	0.498	(0.118)	< 0.0001
1-pentanol	0.079	(0.225)	NS			NS	0.338	(0.299)	0.1	1.14	(0.275)	< 0.0001
2-methyl 1-propanol	3.95	(0.373)	< 0.0001			NS	0.542	(0.640)	NS	3.34	(0.796)	< 0.0001

[1] Parameter estimates for male subjects; female subjects were the reference group (i.e., Est = 0)
$\hat{\sigma}_b^2$; estimated between-subject variance
$\hat{\sigma}_w^2$; estimated within-subject variance
NS; "not significant"

sistent for other exposure or treatment profiles. This allows us to model different scenarios to assess internal dose and metabolism.

For example, consider that we have gathered a series of breath samples pre-, during-, and post-exposure to a well-defined concentration of a specific compound.[12–14] There are a number of ways of interpreting such data using classical pharmacokinetic principles. Researchers including Lance Wallace from US Environmental Protection Agency (EPA), Clifford Weisel from Rutgers University, and James Raymer from Research Triangle Institute constructed closed form solutions from dynamic measurements to assess the time constants and internal (hypothetical) compartments of the human body such as blood and lymphatic fluids, highly perfused tissues including organs and muscle, poorly perfused tissues including tendons, bone marrow, and connective tissues, and finally adipose tissue.[15–17] These models were based on simple exponential uptake and decay behaviors with different time constants assigned empirically to the hypothetical compartments. This provided important information regarding the internal distribution of specific compounds, but did not provide organ specific dose, nor did it allow generalizing to random intermittent exposures.

Subsequently, new methods were developed for interpreting observational kinetic data and estimating response to input functions for the general case. These included very complex physiologically based pharmacokinetic models (PBPK) that require a great deal of specific data, and also prior knowledge about organ dose kinetics, partition functions, and diffusion parameters that are generally only available from invasive animal studies.[18,19] Although PBPK models are powerful tools, they are beyond the scope for this discussion.

A hybrid method was developed by Pleil et al.,[12] wherein the closed form solution was replaced with an iterative solution. This approach requires a starter set of empirical data linked precisely to well-characterized exposure data. Typically, this is generated in a human exposure chamber.[2,20] The methodology is based on conversion of differential equations to difference equations that can then be easily evaluated in standard spread sheet software as follows:

In the simplest form, consider that one has concentration measurements of the central compartment (blood), $C_{\text{blood}}(t)$, and matching environmental air measurements, $C_{\text{air}}(t)$ for some exogenous compound. One further assumes some stable (1^{st} order) rate constant for inhalation uptake (k_u) and some aggregate 1^{st} order elimination rate constant k_e in 1/time units.

Under the assumption of linear kinetics,

$$dC_{\text{blood}}(t)/dt = k_u C_{\text{air}}(t) - k_e C_{\text{blood}}(t).$$

For a well controlled air concentration profile (e.g. $C_{\text{air}}(0) = 0$ and $C_{\text{air}}(t > 0) = C_0$), this differential equation is easily solved with the closed form solution:

$$C_{\text{blood}}(t) = C_0(k_u/k_e)[1 - \exp(-k_e t)]$$

when the boundary conditions are applied. This function can then be fitted to the real data of the proposed step function exposure as could be generated in an environmental chamber, and the rate constants k_u and k_e can be estimated. Regrettably, such closed form solutions are only possible in the simplest conceptual models. In general, there are multiple compartments (blood, highly perfused tissues, poorly perfused tissues, adipose (fatty) tissues, connective tissues, etc.) that all have different capacities and time constants. Each compartment may further exhibit different elimination pathways and time constants as well. As such, the only way to model behavior is via iterative computational methods. We have developed a difference equation approach to compute biomarker concentrations from essentially any complex system given initial conditions, some kinetic data, and a reasonable conceptual model of the compartmental structure.

Using the simple example above, one can rewrite the differential equation as a difference equation in the form of:

$$\Delta C_{\text{blood}}(t)/\Delta t = k_u C_{\text{air}}(t) - k_e C_{\text{blood}}(t)$$

which can be expanded for any arbitrary time increment as:

$$[C_{\text{blood}}(t + \Delta t) - C_{\text{blood}}(t)]/[(t + \Delta t) - t] = k_u C_{\text{air}}(t) - k_e C_{\text{blood}}(t)$$

and subsequently rearranged as:

$$C_{\text{blood}}(t + \Delta t) = C_{\text{blood}}(t) + [k_u C_{\text{air}}(t) - k_e C_{\text{blood}}(t)]\Delta t$$

which is a form wherein each future concentration can be calculated from the previous point for any complex exposure function $C_{\text{air}}(t)$ as long as the rate constants and initial conditions are known. This iterative procedure is successful for more complex models for which we can draw a conceptual diagram because we never need to deduce a closed form solution. Although beyond the scope of this discussion, Pleil[21] has presented a detailed discussion and application of these techniques for blood and breath biomarkers, multiple compartments, and human metabolites.

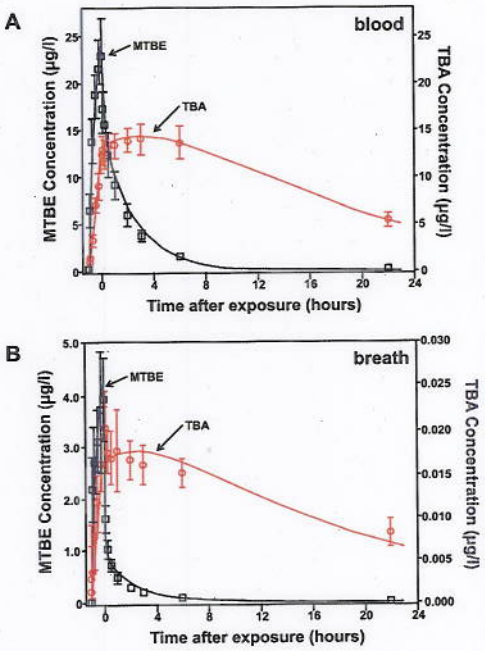

Figure 4. Observed concentrations and empirical models of MTBE and TBA biomarkers in (A) human blood, and (B) human exhaled breath resulting from a 1hr exposure to 3 ppmv MTBE. Note that the phase-1 metabolite response of TBA is damped with half-life of ~ 12 hrs in contrast to the native compound MTBE with half-life of ~ 1 hr.

Figures 4A and 4B show applications of this procedure to time-dependent blood and breath data for methyl tertiary butyl ether (MTBE) chamber exposures and the resulting phase-1 human metabolite tertiary butyl alcohol (TBA). Note that breath data have the same kinetic behavior as the blood data and that the TBA metabolite exhibits biological damping (delayed response) making it a longer time-frame marker for MTBE exposure.

Once the rate constant parameters for the known chamber exposures are calculated, the model can then be applied to any arbitrary exposure input function under the assumption that the concentration lie within the regime of linear kinetics. For environmental levels, this is almost always the case. As such, a solid longitudinal data set from known (controlled) expo-

sures with precise empirical biomonitoring data can be used to implement a generalized model for any exposure profile.

3. Conclusions and recommendations

The premise of this chapter is simple: the value of breath analysis does not end with making a series of measurements. It is critically important to understand the underlying structure of the data, and then to apply the appropriate mathematical and statistical interpretations to be able to use the data to actually answer a particular question. Certainly, the level of detail and the exact techniques are dependent on the available data and meta-data (as well as the question). If one only wants to know what kind of compounds may be found in breath using a chosen sampling and analytical scheme, then simply listing a set of chemicals and their means and ranges may be sufficient. If one wants to determine how fast an anesthetic is eliminated after a surgery or an environmental compound is metabolized, then a simple time series analysis based on classical pharmacokinetics (rate constant) analysis will suffice. If, however, one wants to use breath analysis for discerning the reasons for a particular adverse health outcome, then a much more complex analysis that includes host factor data, suites of compounds, repeated measurements, and a sufficiently large subject pool is required. In short, the data tell a story only if the mathematics and statistics are sound, and also if the sampling design is suited to actually answer the posed question in the first place.

Acknowledgements

The authors thank Terence Risby from Johns Hopkins University, Baltimore, MD, USA, Anton Amann from the Austrian Academy of Sciences, Innsbruck, Austria, Michael Phillips from Menssana Research Inc., Newark, NJ, USA, Wolfram Miekisch from Rostock University Hospital, Rostock, Germany, and Jens Herbig from Ionimed Analytik, Innsbruck Austria for many insightful discussion regarding the use of human breath biomonitoring data. We also thank Linda Sheldon, Tzipporah Kormos, Karen Oliver, and Myriam Medina-Vera of US EPA, and Matthew Stiegel from University of North Carolina, Chapel Hill for their expert advice.

Disclaimer

The United States Environmental Protection Agency through its Office of Research and Development funded and managed the research described

here. This research has been subjected to Agency review and approved for publication.

References

1. Hubbard HF, Sobus JR, Pleil JD, Madden MC, and Tabucchi S. Application of novel method to measure endogenous VOCs in exhaled breath condensate before and after exposure to diesel exhaust. *J Chromatogr B Analyt Technol Biomed Life Sci* 2009; **877**: 3652–8.
2. Sobus JR, Pleil JD, Madden MC, Funk WE, Hubbard HF, and Rappaport SM. Identification of surrogate measures of diesel exhaust exposure in a controlled chamber study. *Environ Sci Technol* 2008; **42**: 8822–8.
3. Pleil JD, Stiegel MA, Sobus JR, Liu Q, and Madden MC. Observing the human exposome as reflected in breath biomarkers: heat map data interpretation for environmental and intelligence research. *J Breath Res* 2011; **5**: 037104.
4. Burmaster DE and Murray DM. A trivariate distribution for the height, weight, and fat of adult men. *Risk Anal* 1998; **18**: 385–9.
5. Pleil JD and Sheldon LS. Adapting concepts from systems biology to develop systems exposure event networks for exposure science research. *Biomarkers* 2011; **16**: 99–105.
6. Pleil JD and Lorber MN. Relative congener scaling of Polychlorinated dibenzo-p-dioxins and dibenzofurans to estimate building fire contributions in air, surface wipes, and dust samples. *Environ Sci Technol* 2007; **41**: 7286–93.
7. Pleil JD, Stiegel MA, and Sobus JR. Breath biomarkers in environmental health science: exploring patterns in the human exposome. *J Breath Res* 2011; **5**: 046005.
8. Rappaport SM and Kupper LL, *Quantitative exposure assessment.* 2008, Stephen Rappaport: El Cerrito, California, USA.
9. Singer JD. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *J Educ Behav Stat* 1998; **23**: 323–55.
10. Sobus JR, McClean MD, Herrick RF, Waidyanatha S, Nylander-French LA, Kupper LL, and Rappaport SM. Comparing urinary biomarkers of airborne and dermal exposure to polycyclic aromatic compounds in asphalt-exposed workers. *Ann Occup Hyg* 2009; **53**: 561–71.
11. Sobus JR, Pleil JD, McClean MD, Herrick RF, and Rappaport SM. Biomarker variance component estimation for exposure surrogate selection and toxicokinetic inference. *Toxicol Lett* 2010; **199**: 247–53.
12. Pleil JD, Kim D, Prah JD, and Rappaport SM. Exposure reconstruction for reducing uncertainty in risk assessment: example using MTBE biomarkers and a simple pharmacokinetic model. *Biomarkers* 2007; **12**: 331–48.
13. Pleil JD and Lindstrom AB. Exhaled human breath measurement method for assessing exposure to halogenated volatile organic compounds. *Clin Chem* 1997; **43**: 723–30.

14. Pleil JD, Fisher JW, and Lindstrom AB. Trichloroethene levels in human blood and exhaled breath from controlled inhalation exposure. *Environ Health Perspect* 1998; **106**: 573–80.
15. Wallace L, Pellizzari E, and Gordon S. A linear model relating breath concentrations to environmental exposures: application to a chamber study of four volunteers exposed to volatile organic chemicals. *J Expo Anal Environ Epidemiol* 1993; **3**: 75–102.
16. Weisel CP, Kim H, Haltmeier P, and Klotz JB. Exposure estimates to disinfection by-products of chlorinated drinking water. *Environ Health Perspect* 1999; **107**: 103–10.
17. Raymer JH, Pellizzari ED, Thomas KW, and Cooper SD. Elimination of volatile organic compounds in breath after exposure to occupational and environmental microenvironments. *J Expo Anal Environ Epidemiol* 1991; **1**: 439–51.
18. Kim D, Andersen ME, Pleil JD, Nylander-French LA, and Prah JD. Refined PBPK model of aggregate exposure to methyl tertiary-butyl ether. *Toxicol Lett* 2007; **169**: 222–35.
19. Kim D and Nylander-French LA. Physiologically based toxicokinetic models and their application in human exposure and internal dose assessment. *Exs* 2009; **99**: 37–55.
20. Prah J, Ashley D, Blount B, Case M, Leavens T, Pleil J, and Cardinali F. Dermal, oral, and inhalation pharmacokinetics of methyl tertiary butyl ether (MTBE) in human volunteers. *Toxicol Sci* 2004; **77**: 195–205.
21. Pleil JD. Influence of systems biology response and environmental exposure level on between-subject variability in breath and blood biomarkers. *Biomarkers* 2009; **14**: 560–71.