Estimating lifetime risk from spot biomarker data and intra-class correlation coefficients (ICC)

Joachim D. Pleil and Jon R. Sobus
Human Exposure and Atmospheric Sciences Division, NERL/ORD
US Environmental Protection Agency (EPA)
Research Triangle Park, NC 27711

ABSTRACT

Human biomarker measurements in tissues including blood, breath, and urine can serve as efficient surrogates for environmental monitoring because a single biological sample integrates personal exposure across all environmental media and uptake pathways. However, biomarkers represent a "snapshot" in time, and risk assessment is generally based on long-term averages. In this article, we propose a statistical approach for estimating long-term average exposures from distributions of spot biomarker measurements using intra-class correlations (based on measurement variance components) from the literature. This methodology is developed and demonstrated using a log-normally distributed data set of urinary oh-pyrene taken from our own studies. The calculations are generalized for any biomarker data set of spot measures such as those from the National Health and Nutrition Evaluation Studies (NHANES) requiring only spreadsheet calculations. We develop a three-tiered approach (depending on the availability of meta-data) for converting any collection of spot biomarkers into an estimated distribution of individual means that can then be compared to a biologically relevant risk level. Examples from a Microsoft Excel® based spreadsheet for calculating estimates of the proportion of the population exceeding a given biomonitoring equivalent level are provided as an appendix.

Introduction

Health risk assessment is the mechanism that shapes public policy with respect to controlling and regulating chemical compounds and biological stressors in the environment. The World Health Organization (Prüss-Üstün and Corvalán, 2006) has estimated that:

> *"An estimated 24% of the global disease burden and 23% of all deaths can be attributed to environmental factors."*

A large proportion of this environmental risk is attributable to waterborne disease vectors resulting in diarrhea and other infectious diseases. Other major contributors are inhalation exposures to particulate matter and gas-phase chemicals, as well as other chemicals via dermal and ingestion exposures, which are the focus of this article. Such exogenous chemical exposures contribute to respiratory infections, and chronic diseases such as COPD, ischemic heart disease, asthma, lung and other cancers. Rappaport and Smith have estimated that 70% to 90% of all chronic diseases (autoimmune, cancer, cardiovascular, etc.) are attributable to *"…differences in environments.*" (Rappaport and Smith, 2010).

There are a variety of techniques in current practice wherein environmental measurements are interpreted to estimate population-based risks of adverse health outcome (Brunekreef, 2008; Nachman et al., 2011; Pleil et al., 2012a). Of these, cancer outcomes are probably the best understood in that epidemiological studies have shown associations between particular environmental exposures and specific cancer types (e.g., Chen et al., 1992; Perera, 1997; Pope et al., 2002; Steffen et al., 2004; Wolff and Weston, 1997). Recent work has advanced such risk assessments to incorporate human biomarker measurements as a bridge between external exposures and ultimate disease incidence. U.S. Environmental Protection Agency (EPA) has published a biomonitoring framework that integrates measurements with models (see Figure 1) (Sobus et al., 2011; Tan et al., 2012a, Tan et al., 2012b). Here, the progression of empirical measurements from environmental media, to internal dose, to biologically relevant dose, to early effect and disease outcome is shown in parallel with interconnections to stochastic and kinetic models. Both measurements and models are equally important; the measurements provide the "ground truth" and the models provide the relationships and estimates where measurements are not possible or available. We note that the ultimate goal for using environmental and biomarker data in this way is to mitigate lifetime risk from long-term latency diseases; this requires population-based inferences using contemporary data and modeling conventions. Extending such information to reflect a whole lifetime is certainly dependent on many different assumptions and their respective error bounds, but is necessary, as current risk estimates require such information.

From this conceptual model, we could assign health risk more directly by using measurements of individual dose or body burden as a replacement for numerous environmental measurements. We note that the ensuing discussions are geared towards health risk from long-term or chronic exposures at environmental (low-levels) and not acute (high-level) exposures that could have an immediate health outcome such as those from carbon monoxide,

cyanide, pesticide, arsenic, or lead poisoning, among others. The advantages of using biomarkers of all types of exposures and effects are manifold.

- Biomarkers avoid the "ecologic fallacy" wherein a common air, water, or food measurement is assigned to all individuals in a group.
- Biomarkers integrate across all exposure pathways (e.g. inhalation, ingestion, dermal contact) simplifying sampling requirements.
- Biomarkers are "closer" to the biologically relevant dose because differences in activity patterns (e.g. hobbies, work), physiological parameters (e.g. breathing rate, water consumption) and metabolism are accounted for at the individual level.

The correspondence between environmental exposures and internal human biomarkers can be established using certain models and/or paired empirical measurements (Angerer et al., 2011; Aylward et al., 2012b; Hays and Aylward, 2012; Hays et al., 2012; Pleil et al., 1998; Pleil et al., 2007). These kinds of studies establish parameters for estimating direct links between accepted risk "slope" factors for environmental concentrations and individual biological measurements that exploit the advantages listed above. The one problem that this approach does not solve, however, is the interpretation of measurement variance. Cancer risk models rely on the concept of average daily rate of intake (ADRI) that assumes that the lifetime risk for disease incidence is proportional to average exposure and thus relatively unaffected by short-term fluctuations (Crawford-Brown, 1997; Pleil et al., 2004). However, single snapshot measurements of internal biomarker compounds are not likely representative of a lifetime average. Not only do they vary between individuals, they can also vary temporally within individuals due to changes in diet, health state, mobility, location, etc.(Lin et al., 2005; Pleil, 2009; Sexton et al., 2005; Sexton and Ryan, 2012; Sobus et al., 2010).

Large databases of biomarker measurements from National surveys (such as the US National Health and Nutrition Examination Survey [NHANES] and the German Environmental Survey [GerES]) are attractive for risk assessment efforts because they include thousands of individuals and hundreds of empirical measurements per person, but are difficult to interpret because they lack environmental exposure data and do not make repeat measurements of biomarkers for individual participants. As such, the data are useful for providing a health and nutrition based "report card" for tracking temporal trends across many years, but cannot be directly linked from environmental exposures to their attendant individual risks.

Given a distribution of a particular biomarker across a study population, the best we can do is to use the summary statistics to calculate a mean risk for the population and possibly bracket this risk with total variance estimates. What cannot be determined is what percentage of the population exceeds a particular risk level. This can only be assessed with detailed knowledge of the relative contribution of the within-subject and between-subject variance components to the total observed (cross-sectional) variance (Rappaport and Kupper, 2008; Tornero-Velez et al., 1997). Researchers have proposed estimating such risk exceedances using pharmacokinetic models based on absorption, distribution, metabolism, and elimination (ADME) kinetics

(Aylward et al., 2012a). Although such approaches are valuable, they also require detailed information for specific chemicals and are difficult to generalize.

The methods developed in this article are motivated by scientific concerns regarding the environmental impact on health and how to use existing data to assess and mitigate risk. Some common questions are:

- How can we measure population risks?
- What are the risk levels of concern?
- How can public policy reduce risks of concern?
- How many individuals in a group have risks above a certain defined level?
- What is my risk in comparison to the average risk?

The first two questions regarding overall public health risks are addressed with statistical methods based on measurements of environmental concentrations of contaminants coupled to in vitro experiments with cell lines and dose response measurements in animal models (Pleil, 2012; Pleil et al., 2012b). The preponderance of evidence for specific compounds, or groups of compounds with similar mode of action, is used to develop risk factors or "slope factors" that can be used to estimate population based risk from environmental concentrations (Andersen et al., 2010; Ankley et al., 2010).

The third question concerning public policy is generally addressed with broad-based advice or rules including emissions standards for factories, implementation of catalytic converters for automobiles, regulations against indoor smoking, reducing sulfur content in diesel fuels, etc. These types of efforts shift the average environmental exposures to lower levels thus reducing the population based ADRI; they do not necessarily affect the shape of the distributions of the concentrations.

The remaining two questions about what happens at the individual level are currently largely unanswered. We propose specific methods for filling this void and interpreting existing data to assess distribution of risk. We explore a simple statistical approach wherein available spot biomarker measurements from large databases are interpreted to estimate environmental health risk without implementation of complex models. We develop a general approach for calculating "worst case" population risk, and then use underlying biomarker distributions to bound the range of risk on an individual basis as a function of intra-class correlation coefficients (ICC) defined as between-subject variance divided by total variance. We also provide examples for refining risk estimates from NHANES data directly when some ancillary distribution data is available from other (typically smaller and more focused) studies. The ability to augment population based risk estimates with insight into the distribution of individual risks will ultimately demonstrate the value of any public policy effort.

Methods

Standard environmental methods:

Risk assessments are generally made at the population level using environmental measurements at the macro scale (e.g. citywide); that is, everyone in a particular group is assigned the same mean risk with some calculated confidence intervals (for example, see EPA's Guidance for using the National Ambient Air Quality Standards for estimating risk: http://www.epa.gov/ttnnaaqs/standards/no2so2sec/cr_rea.html). When internal biomarker measurements are available, the error associated with uptake calculations and source identification could be reduced, but risk estimates are still based on mean values coupled with slope factors. Currently, there are no regulatory processes informing risk at an individual level.

Studies have implemented stratified approaches wherein macro scale measurements (city wide) are supplemented with meso scale (neighborhood wide) and micro scale (residential) environmental measurements (Bereznicki et al., 2012; Rodes et al., 2010). These methods allow us to assess the spatial variances in exposure levels and achieve an understanding of the exposure concentration "surface". The ultimate goal is to determine if a central measurement (e.g. on top of the post-office) is relevant for assigning the exposure to someone living two miles away.

To reach the individual level requires an individual measurement. Some historical studies had implemented the concept of "personal dosimetry" to further stratify the micro-scale environment (Frazier et al., 2009; Williams et al., 2008). These were primarily focused on inhalation exposures and made measurements directly in the breathing zone of individual subjects wherein they carried small sampling devices attached to their clothing near their mouths. In addition to the examples quoted for airborne exposure, other studies have made individual "food basket" assessments wherein duplicate diet items were collected and analyzed for specific compounds such as pesticides (Curl et al., 2003; Lu et al., 2010; Vogt et al., 2012).

Using a combination of data and models, the risk assessment community makes representative guidelines based on estimates of ADRI (ng/(kg x day), a calculated slope factor (cancer incidence/(ng/kg x day) x 70yrs x 365 days/yr) and some acceptable risk level for comparison, typically 10^{-5} or 10^{-6} (referring to cancer incidences of 1/100,000, or 1/1,000,000 people, respectively). U.S. EPA maintains data in support of human health risk assessment through a program named Integrated Risk Information System (IRIS) that provides such slope factor information. (e.g. http://epa.gov/iris/, http://www.epa.gov/cancerguidelines/).

Biomarker methods:

The next step in environmental methods development includes direct biological samples from individuals. For population-based studies (typically for many subjects with 1 sample per subject), this approach provides a direct "exposure snapshot" of individuals within the sampled group. Such biomarker data are not yet implemented for risk assessment in that risk slope factors are developed from environmental measurements and so any form of biomarker measurement requires an environmental exposure reconstruction. Recently, researchers have proposed the use of "biomonitoring equivalents" (BE) wherein a biomarker measurement is used as a surrogate for external exposure (Angerer et al., 2011; Aylward et al., 2012b; Becker et al., 2012; Hays and Aylward, 2012). Such BE values are calculated using a broad spectrum of available epidemiological, animal, and *in vitro* studies and determining reasonable values of BE's is beyond the scope of this paper. For statistical purposes here, however, we define that a specific BE for a specific compound is designated to represent an onboard level of a specific biomarker can be related back to a mean exposure level of concern. We implement the BE parameter concept as a choice for the risk assessor to make, and use it to show how the BE is related to the underlying calculated distributions of the biomarker.

A second caveat to using biomarkers concerns the representativeness of a single measurement with respect to a lifetime exposure. The nature of a biomarker measurement, whether of an exogenous compound or a chemical metabolite, is that human systems biology may affect the short-term variability in unknown ways. Furthermore, external variability driven by human activities and interaction with the micro-environment can introduce short-term changes independent of systems biology (e.g. refueling a car, using a cleaning product, spraying for insects). Unlike environmental measurements that are typically averaged over 24-hrs and subsequently repeated daily over many years (e.g. National Ambient Air Quality System (NAAQS) http://www.epa.gov/air/criteria.htmlref), the chemical biomarker reflects "now", and even longer lived metabolites in urine, for example, reflect the past few hours, or the previous day, at most. As such, the solution to extending a risk estimate from snapshot measures to lifetime averages is to have a full understanding of the variance components surrounding any individual measurement.

Variance components of biomarkers:

Biomarker measurements made from "n" individuals typically result in a lognormal distribution; any single value reflects the current status of that particular individual. The variability associated with any single measurement can have extreme values that range from zero to the whole distribution depending upon the nature of the compound, the exposure profiles, and the classical (ADME) pharmacokinetics of the individual. For example, if all repeat measurements from a specific participant are the same over time, then the initial single measurement represents this person's long-term exposure. If, on the other hand, repeated measurements from a single participant could span the range of the total distribution, then this person's long-term exposure is proportional to the mean of all of the initial measurements.

Falling at either of these extremes is unlikely; in general, repeat measures are correlated to some extent even when taken at longer intervening times. The statistical parameter that describes this condition is the intra-class correlation coefficient (ICC) defined as the between-subject variance (σ_b^2) divided by the sum of the between-subject and within-subject variance: $(\sigma_b^2)/(\sigma_b^2 + \sigma_w^2)$. ICC can range from 0 to 1; when ICC is close to 0, we expect that any repeated measure from an individual can take any value across the distribution and when ICC is close to 1, repeated measurements from a particular individual are expected to stay the same.

If biomarker levels are assigned to reflect an individual risk using a linear multiplier (slope-factor), then the population risk is proportional to the central tendency of all measurements, regardless of ICC. However, if we want to establish how many individuals in a particular population have a lifetime risk estimate that falls beyond a certain criterion of concern, then a higher ICC is statistically linked to a larger number of persons in that exceedance category. One could thus establish a bracket for number of exceedances for any population of measurements and established risk calculations, and eventually calibrate the correct level within that bracket based on additional knowledge of ICC.

Construction of an example data set:

The methods described above are demonstrated using a log-normally distributed data set based on urinary measurements of polycyclic aromatic hydrocarbons (PAHs) metabolites from our own work where we have specific statistical insight (Sobus et al., 2009a; Sobus et al., 2009b; Sobus et al., 2010). Specifically, we chose 1-OH-pyrene from the suite of available compounds for this demonstration. Although the original data set was unbalanced (comprised of 215 measurements from 30 participants with 2 to 10 samples per person), it is realistic, and known to be log-normally distributed with an ICC of 0.507 as calculated in log-space. For demonstration purposes, we use these data as if they were independent measurements, and have modified them slightly by adding a few imputed values (to round up to n=220). We calculated summary statistics with geometric mean GM = 1155 ng/L and geometric standard deviation GSD = 2.84 ng/L. For the remainder of this discussion, these two parameter values are designated as the "global" spot measures distribution and carry a subscripted "g" (e.g. GM_g and GSD_g).

Construction of hypothetical "within-subject" distributions:

To illustrate the effect of ICC on the eventual risk distribution, we have used the initial data set and have created hypothetical repeat measures recalculated and then drawn from the same global distribution of values. We then randomly stratified the repeat measures to achieve a range of ICC's for investigation by clustering higher values together for some subjects, middling values for some, and lower values together for others etc. We used an "overlapping group" strategy wherein values were sorted by 2, 5, 11, 22, and 110 subjects at a time, in addition to the totally random (ICC=0), and totally ordered (ICC=1) scenarios. For example, for the grouping strategy of 11 subjects per group, we start with 220 discreet values and assign the lowest 22 values at random to the first group of 11 subjects, the next lowest 22 values to the

second group, etc., until the highest 22 values are left to be assigned at random to the last group. This was repeated "m" times to achieve different hypothetical repeat measures. The smaller we choose the subject per group assignment, the smaller the eventual ICC calculation becomes. The hypothetical repeated measurements were created in Excel® 2010 (Microsoft, Redmond WA) using "lognorm.dist(true)" and "rand( )" functions to achieve random stratified values within persons that maintained the global summary statistics of GM_g = 1155 and GSD_g = 2.84 ng/L. We caution that earlier versions of Excel® are lacking some of this functionality. As such, the overall (global) summary data (mean, standard deviation, total variance, distribution, etc.) remain the same, the only differences among the new data-sets is how the variance components are distributed. As a result, we created different ICCs from the same basis set of measurements to assess the influence of relative variance components. We note that ICC = 0 corresponds to the "best case" risk scenario and ICC = 1 corresponds to the "worst case" risk scenario.

Calibrating ICCs with respective risk distributions:

To make these concepts useful for estimating means distributions for any generic data set, we needed to first link externally determined ICC's with distributions of individual means. We developed a series of hypothetical data sets based on the global data seen in Figure 2 and empirically calculated ICC's for different numbers (2, 5, and 10) of measurements per individual (m). These results were investigated to achieve a general calibration technique realizing that the links between ICC and means distribution were likely dependent on both m and the value of ICC. Given such calculated values, we explored the relationships among log-normally distributed means values and attendant ICC's and developed a generic technique for establishing this relationship for any combination of parameters.

Estimating proportion of individuals above a given BE:

Once a means distribution is established from a set of spot measurement data, the final step in the method is to calculate how much of the curve area is beyond some BE value of interest and calculate the ratio of the tail area to the total area. Because we have established a means frequency distribution, this ratio is now a measure of the proportion of the population that is expected to exceed the BE for a lifetime exposure. Under the assumption that the means distributions are lognormal, these areas were calculated using the "lognorm.dist (a2,a3,a4,true)" function in Excel®; these could also be calculated with other commercially available software packages.

Results

Example data set

Figure 2 shows the frequency distribution of 220 biomarker measurements of urinary oh-pyrene composited from earlier work as discussed above. These data represent a generalized example of such measurements for demonstration purposes and include a wide range as could be expected from a randomized study wherein both incidental environmental and occupational exposures are possible. For this article, we treated the values as independent spot measures, each from a different individual. We defined μ = mean, and σ = standard deviation of the logged data, respectively, and geometric mean GM = exp(μ) and geometric standard deviation GSD = exp(σ) in normal space.

Each spot measurement in Figure 2 could represent that individual's mean value for lifetime exposure ("worst case"), or a random value of current exposure from the global distribution ("best case"), or some stratified value in between. To demonstrate the effect of internal stratification, we used the GSD of the original data to serve as the GSD_w of the "worst case", and we calculated the standard error of the mean (SEM) of the logged spot measures: SEM = $\sigma/\sqrt{220}$, and converted it to GSD_b = exp(SEM) for the best case distribution. Figure 3 shows the extreme distribution curves for n = 220 measurements with common (global) GM_g = 1155 ng/L and $GSD_w = GSD_g$ = 2.84 or GSD_b = 1.023 ng/L, respectively. If we choose BE = 4400 ng/L as a departure point of interest, we see that for the worst case scenario, 10% of the population has a biomarker mean value greater that BE, whereas for the best case distribution, 0% of the population means exceed the value.

Within subject distributions

The literature abounds with biomonitoring studies wherein individuals are sampled repeatedly. These generally use smaller total numbers, but can have anywhere from 2 to 10 or more measurements/person spread out over different time frames. For the ensuing demonstration, we generated a series of hypothetical repeat measurements to align with our basis data set for oh-pyrene urinary biomarkers and performed random stratification to achieve different ICC's using Excel®. Subsequently, we calculated ICC for each set of stratified data using PROC MIXED REML estimates in SAS (Cary, NC). Although we made all calculations based on the full data set of 220 spot measurements, we randomly selected a subset of 25 out of the 220 hypothetical subjects from this set for demonstration purposes.

Figure 4 shows box and whiskers plots for 25 hypothetical subjects with 10 repeat measurements each, all drawn from different re-arrangements of the global data and the respective resulting ICC's. The internal structure of biomarker data and the resulting variability of the subject means are now described by a range of ICC's. We note that when ICC is closer to 0, the repeat measures distributions within subjects are large and thus overlap with each other, and the mean values across subjects are fairly similar. As ICC approaches 1, the within subject variability decreases and the subject means become more distinct from each other. We point

out that all individual measurements within each graph come from the same global distribution: GM_g = 1155, GSD_g = 2.84. The only parameter that changes is ICC, which indicates how these measurements are distributed among the subjects.
Figure 4. In each case, 25 subjects with 10 measurements/subject were selected at random from data stratified based on ICC outcome. As ICC increases, the within subject scatter decreases and the between subject means distinction increases. For each graph, the overall distribution of spot measures remains the same at GM_g = 1155 and GSD_g = 2.84.

Calibrating ICCs to risk estimates:

Above (in Figure 4), we demonstrated that the distinction among subject mean values increases with higher ICC. Furthermore, in the risk assessment models, we recall that the eventual health outcome is linked to a subject mean value, not a spot measurement. As such, the next step is to link ICC directly with a distribution "spread" parameter that will allow us to better estimate risk by calculating the number of subjects with mean biomarker levels beyond some BE of interest.

We recognize that the quality of our empirical ICC calculations depends primarily upon "m", the number of measurements per subject, and to lesser extent, on the total number "n" of subjects studied. Certainly estimates based on m=10 are expected to provide better ICC estimates for a population than m=2. We explored the relationships among ICC, m, and the geometric distribution of the means using the approach described above. We made the same calculations for n = 220, m=2, 5, and 10 to discover the strength of the effect of increasing m. Figure 5 shows empirical results for these relationships; we see that we can generate a straight-line response between ICC and the means distribution "spread" parameter "GSD/GM_g". We note that GM_g is a constant for the overall distribution, only GSD changes as a function of ICC, but we chose this display method to make the x-axis dimensionless to remove any confusion from different choices of measurement units. This approach also facilitates direct comparisons among different compounds.

From this initial experiment, we found that the x-intercept is a critical parameter for establishing the slope of the curves as a function of m. Upon detailed statistical analysis, we constructed a generic function for calculating ICC vs. GSD response.

We define the important parameters as follows:

σ = standard deviation of logged data
μ = mean of logged data
GSD_g = $\exp(\sigma)$ = "global" geometric standard deviation of the initial data set
GM_g = $\exp(\mu)$ = "global" geometric mean of the initial data set
X_{min} = exp{ln[GSDg]/sqrt(m)}/GMg when ICC = 0
X_{max} = GSD_g/GM_g when ICC = 1
m = number of measurements per person

Based on the graphs above, the means distribution GSD for any set of spot measurements as a function of ICC and m is defined by:

$$GSD = GM_g \times \{ICC(m) \times (Xmax - Xmin) + Xmin\} \qquad \text{eq. 1}$$

We can now generate the distribution of the means for any set of spot measures under the assumption that they are normally distributed based on the specifically calculated GSD for the particular ICC parameters and GM_g, the global geometric mean of the biomarker spot measures.

Calculating the exceedance beyond a BE of interest:

Given the distribution of the subject lifetime means, we can now calculate the exceedance beyond some BE of interest. The empirical equations for this type of assessment are somewhat complex, however, various statistical software packages have built-in functions for providing a direct result. In Excel®, for example, the function:

$$Area\% = 100 \times [1 - lognorm.dist(BE, \ln(GM_g), \ln(GSD), true)] \qquad \text{eq. 2}$$

returns the percent of the total of the area under the curve beyond BE of the lognormal density function described by the logged mean and logged standard deviation of the means distribution. Figure 6 shows some example curves and quantitative results of such calculations.

For the preceding demonstration example, we chose m=10 as a "moderately accurate" number of measurement for calculating ICC's. As mentioned above, however, real-world study data vary widely; we have seen a range of 2 to more than 70 measurements per subject. In Figure 5 above, we showed that the increasing confidence in the ICC from higher "m" allows us to make a better estimate as to the exceedance of the population beyond BE. In Table 1, we show the relative change in estimates dependent on repeat measures.

Table 1. Relationships among ICC's and repeat measures for estimating the % area beyond BE = 4400 ng/L for urinary oh-pyrene biomarker level from the example study.

	m = # of measurements/subject					
ICC	**~1000**	**50**	**10**	**5**	**2**	**1**
0.00	**0.00**	**0.00**	**0.00**	**0.22**	**2.98**	**na**
0.25	**0.04**	**0.16**	**0.96**	**1.94**	**4.82**	**na**
0.50	**2.14**	**2.60**	**3.85**	**4.68**	**6.67**	**na**
0.75	**6.14**	**6.39**	**7.13**	**7.49**	**8.41**	**na**
1.00	**10.00**	**10.00**	**10.00**	**10.00**	**10.00**	**10.00**

These results show that there is indeed value in greater number of measurements per subject for estimating the lower bound of exceedance beyond a BE of interest. The last column for m = 1 demonstrates the default case when there are no repeat measures at all for calculating an ICC. Here, the best we can do is to assume that each measurement represents that subject's mean value and so we assign ICC = 1. We note that for this article we have assumed that BE has a fixed value imposed externally; in reality, estimates of BE may exhibit their own statistical distributions which must be accounted for in the long run using epidemiological and other human studies data.

Discussion

Evaluating the risk from environmental chemicals based on spot biomarker measurements requires some form of projection to long-term average values for individual subjects in a population. Probably the best such estimates would come from detailed knowledge of the pharmacokinetics (ADME) coupled with exposure measurements. As very few chemicals have sufficient data available to achieve this level of knowledge, we have explored more accessible methodology for developing means distributions. Throughout the development of the mathematical approach in this article, we have found that there are three basic tiers of external data for estimating variance components. The first tier is straightforward: there are no available statistical data for ICC at all and so we assume that each spot measurement represents the subject mean. The second tier uses an approximation of ICC for a specific compound derived from soft data like inference from observational exposures, similarities to other compounds, and biological lifetimes. The third tier is most robust and exploits calculations of ICC with known numbers of repeat measures from other, albeit smaller, focused studies.

Tier 1: Global bounds

In general, large biomarker data sets like the NHANES are cross-sectional without repeat measures; as such, there is no direct method for calculating ICC. Under these circumstances, a bracket approach calculating the means distribution for ICC = 0 and for ICC = 1 are the only choices for establishing population exceedance estimates. As shown above, this defaults to using either the geometric mean (GM_g) value of the spot biomarker measurements as a central tendency for comparison to BE, or using the spot biomarker distribution itself to represent the means distribution. As shown for the oh-pyrene example, the best approach under these circumstances is to set a lower bound near 0 if BE > GM_g, and an upper bound based on the percent area in the tail as calculated from the frequency distribution characterized by the geometric standard deviation (GSD_g) and GM_g of the original spot measurements (eq. 2). We caution that using estimates below the default of ICC = 1 approximation is not statistically defensible in the absence of other data.

Tier 2: Inferred ICC's

For many compounds, estimates of ICC can be made from information in the literature based on the experience of the researchers, similarities with other compounds with robust ICC measurements, or from calculations and models of environmental measurements and human uptake and elimination parameters. These must be treated with some caution, and, in the absence of articulated repeat measurements, the best approach is to assign the most conservative value, m=2 measurements/subject, for making subsequent calculations. We have shown above that this extra piece of information provides some advantage in refining the means distribution summary statistics by allowing a lower geometric standard deviation to be used to calculate exceedance above BE (eqs. 1, 2). Table 1 shows this improvement effect for

the particular example of oh-pyrene, wherein a modest estimate of ICC = 0.5 with m=2 reduces the percent exceedance from the default value of 10% to 6.7%.

Tier 3: Robust ICC's

There are many focused studies for different compounds that have measured within- and between-subject variance components in normal (unremarkably exposed) people. These are the lynchpin for the approach developed in this article in that they provide the additional information for estimating individual means from databases of spot biomarkers. In general, it is difficult to judge the accuracy of published ICC's without access to the raw data; however, the information provided by "m", the number of measurements per subject, is a good indicator for ICC quality. In addition, "m" serves as an important parameter for setting the extremes of the calibration between the means distribution spread (GSD/GM_g) and the ICC according to eq. 1. We have explored the effects of ICC and "m" parameters in assessing the means distributions using log-normal biomarker data. We found that we can calculate the percent exceedances past some given BE value of interest using a straightforward Excel® function. The estimate is refined beyond the default assumption that m = 2 (prescribed in Tier 2), and results in a lower, yet defensible, estimate for the GSD of the means distribution for m >2, and an attendant reduction in percent exceedance.

Implementation of risk estimates

Regardless of the tier level of external information, the procedures developed in this article improve risk estimates based only on spot measurements. This requires the GM and GSD that can readily be calculated from biomarker databases (e.g. NHANES), some external information regarding variance components (ICC and "m"), and a BE value that describes a level of interest. Under the hypothesis that risk can be associated with mean biomarker level, and thus with an estimate of ADRI, we can calculate the mean and SEM for each subject for each hypothetical "within-subject" distribution. The specific procedures are described in the results section.

Appendix 1 shows snapshots of the quantitative approach in Microsoft Excel® spreadsheets, one for the oh-pyrene example and a second for blood-borne benzene distribution for smokers. The ultimate user can input numeric values for the distribution parameters (μ and σ estimated from logged population data, or GM_g and GSD_g estimated from the natural space data), a BE value of interest, an ICC value, and a value for "m". The spreadsheet is set-up to graph the default "worst case" means distribution corresponding to ICC = 1, and overlay the calculated means distribution for the provided ICC and "m" combination. In addition, the spreadsheet calculates the percent exceedance in the population based on the relative area beyond the BE value for both curves. The authors will provide working Excel® files, instructions, and technical assistance upon request.

Population statistics and other cautions

Certainly, the methodologies presented here are based on the underlying observation that biometric data are generally log-normally distributed. Secondly, there is an assumption that whatever external data is applied to glean ICC parameters is expected to be representative of that population of spot measures. As such, we recommend prudence; it is straightforward to confirm the distribution of log transformed spot measures with Kolmogorov-Smirnov or Shapiro-Wilk tests (depending upon availability of particular software). Also, when using external ICC data, it is important to recognize the particular sub-population studied for calculating ICC and how it relates to the broader population. For example, urinary cotinine measurements are often used to infer smoking status in human subjects. As such, the distributions of cotinine in smokers and non-smokers are expected to be different, and so population distribution tests and ICC meta-data are crucial in matching parameters appropriately.

Summary

The current use of biomarker spot measures distributions offer only two extremes for making population risk estimates. At the low end, one assumes that the geometric mean represents the central tendency of all individuals and at the high end one assumes that each spot-measurement represents the long-term mean for that individual. Neither is likely accurate; in reality, the actual means distribution likely lies somewhere between these two bounds. The approaches developed in this article provide a defensible mechanism for estimating the true means distribution (between the two extremes) by exploiting external data about variance components. The calculations are straightforward and shown to be implemented in a standard spreadsheet. Based on the described tiers approach, better risk estimates are possible with less than perfect variance component (ICCs) data; and the estimates improve as the repeat measures data become more robust.

Because ICC is based on variability, we would expect that compounds with demonstrated long biological half-lives such as certain metals or persistent organic compounds (e.g. dioxin/furan congeners) would have ICC values closer to 1. In general then, this work shows that spot measurements of such compounds could be used as an estimate of long-term exposure for the individual. We caution this may not always be true as a large (unknown) bolus from the environment could create a temporary change that would be reflected in a single measurement. As such, half-life could be used as an indicator of ICC, but actual calculation of ICC is a more robust approach.

The next step in implementing distributions of biomarkers and the corresponding ICC's will be the development of further analytical and statistical evaluations of the BE values themselves. For now, BE is treated as a given single value provided externally. We recognize that there are likely variance components of BE that will impact risk assessment at the individual and group

levels that are not yet considered. We expect that a robust BE knowledge base will become an indispensible tool for future biomarker studies in reducing uncertainties in the attendant risk estimates.

The general implementation of these methods will reduce pressure for making broad-based repeat measures in large cross-sectional studies that would be resource intensive. We imagine that even a small subset of repeated biological measurements could improve spot measurement based risk estimates greatly. For example, consider a large study such as NHANES wherein thousands of subjects are sampled every cycle. If just a few hundred subjects were resampled 3- times (or more) in a time frame that minimizes (temporal) autocorrelation, we could quickly estimate the variance components and refine the means distributions. Furthermore, once the ICC's are reasonably established, they will likely remain stable in time and subsequent cycles could reduce the level of repeated measures without loss of confidence in the statistics.

Acknowledgements

The authors would like to thank Stephen Edwards, Peter Egeghy, Martin Phillips, Myriam Medina-Vera, Krista Christensen, Timothy Buckley, Lynn Flowers, and Linda Sheldon from US EPA, Matthew Stiegel from the University of North Carolina, Chapel Hill and Stephen Rappaport from the University of California, Berkeley for valuable insights and discussions. This work was reviewed by the U.S. EPA and approved for publication, but does not necessarily reflect official Agency policy. The authors declare they have no competing financial interests.

References

Andersen, M.E., Al-Zoughool, M., Croteau, M., Westphal, M., Krewski, D., 2010. The future of toxicity testing. Journal of toxicology and environmental health Part B, Critical reviews 13, 163-196.

Angerer, J., Aylward, L.L., Hays, S.M., Heinzow, B., Wilhelm, M., 2011. Human biomonitoring assessment values: approaches and data requirements. International journal of hygiene and environmental health 214, 348-360.

Ankley, G.T., Bennett, R.S., Erickson, R.J., Hoff, D.J., Hornung, M.W., Johnson, R.D., Mount, D.R., Nichols, J.W., Russom, C.L., Schmieder, P.K., Serrrano, J.A., Tietge, J.E., Villeneuve, D.L., 2010. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environmental toxicology and chemistry / SETAC 29, 730-741.

Aylward, L.L., Kirman, C.R., Adgate, J.L., McKenzie, L.M., Hays, S.M., 2012a. Interpreting variability in population biomonitoring data: role of elimination kinetics. Journal of exposure science & environmental epidemiology 22, 398-408.

Aylward, L.L., Kirman, C.R., Schoeny, R., Portier, C.J., Hays, S.M., 2012b. Evaluation of Biomonitoring Data from the CDC National Exposure Report in a Risk Assessment Context: Perspectives across Chemicals. Environmental health perspectives.

Becker, R.A., Hays, S.M., Robison, S., Aylward, L.L., 2012. Development of screening tools for the interpretation of chemical biomonitoring data. Journal of toxicology 2012, 941082.

Bereznicki, S.D., Sobus, J.R., Vette, A.F., Stiegel, M.A., Williams, R.W., 2012. Assessing spatial and temporal variability of VOCs and PM-components in outdoor air during the Detroit Exposure and Aerosol Research Study (DEARS). Atmospheric Environment 61, 159–168.

Brunekreef, B., 2008. Environmental epidemiology and risk assessment. Toxicology letters 180, 118-122.

Chen, C.J., Chen, C.W., Wu, M.M., Kuo, T.L., 1992. Cancer potential in liver, lung, bladder and kidney due to ingested inorganic arsenic in drinking water. British journal of cancer 66, 888-892.

Crawford-Brown, D.J., 1997. Theoretical and mathematical foundations of human health risk analysis: biophysical theory of environmental health science Kluwer Academic Publishers.

Curl, C.L., Fenske, R.A., Elgethun, K., 2003. Organophosphorus pesticide exposure of urban and suburban preschool children with organic and conventional diets. Environmental health perspectives 111, 377-382.

Frazier, E.L., McCurdy, T., Williams, R., Linn, W.S., George, B.J., 2009. Intra- and inter-individual variability in location data for two U.S. health-compromised elderly cohorts. Journal of exposure science & environmental epidemiology 19, 580-592.

Hays, S.M., Aylward, L.L., 2012. Interpreting human biomonitoring data in a public health risk context using Biomonitoring Equivalents. International journal of hygiene and environmental health 215, 145-148.

Hays, S.M., Aylward, L.L., Driver, J., Ross, J., Kirman, C., 2012. 2,4-D exposure and risk assessment: comparison of external dose and biomonitoring based approaches. Regulatory toxicology and pharmacology : RTP 64, 481-489.

Lin, Y.S., Kupper, L.L., Rappaport, S.M., 2005. Air samples versus biomarkers for epidemiology. Occupational and environmental medicine 62, 750-760.

Lu, C., Schenck, F.J., Pearson, M.A., Wong, J.W., 2010. Assessing children's dietary pesticide exposure: direct measurement of pesticide residues in 24-hr duplicate food samples. Environmental health perspectives 118, 1625-1630.

Nachman, K.E., Fox, M.A., Sheehan, M.C., Burke, T.A., Rodricks, J.V., Woodruff, T.J., 2011. Leveraging Epidemiology to Improve Risk Assessment. The Open Epidemiology Journal 4, 3-29.

Perera, F.P., 1997. Environment and cancer: who are susceptible? Science 278, 1068-1073.

Pleil, J.D., 2009. Influence of systems biology response and environmental exposure level on between-subject variability in breath and blood biomarkers. Biomarkers : biochemical indicators of exposure, response, and susceptibility to chemicals 14, 560-571.

Pleil, J.D., 2012. Categorizing biomarkers of the human exposome and developing metrics for assessing environmental sustainability. Journal of toxicology and environmental health Part B, Critical reviews 15, 264-280.

Pleil, J.D., Fisher, J.W., Lindstrom, A.B., 1998. Trichloroethene levels in human blood and exhaled breath from controlled inhalation exposure. Environmental health perspectives 106, 573-580.

Pleil, J.D., Kim, D., Prah, J.D., Rappaport, S.M., 2007. Exposure reconstruction for reducing uncertainty in risk assessment: example using MTBE biomarkers and a simple pharmacokinetic model. Biomarkers : biochemical indicators of exposure, response, and susceptibility to chemicals 12, 331-348.

Pleil, J.D., Sobus, J.R., Sheppard, P.R., Ridenour, G., Witten, M.L., 2012a. Strategies for evaluating the environment-public health interaction of long-term latency disease: the quandary of the inconclusive case-control study. Chemico-biological interactions 196, 68-78.

Pleil, J.D., Vette, A.F., Johnson, B.A., Rappaport, S.M., 2004. Air levels of carcinogenic polycyclic aromatic hydrocarbons after the World Trade Center disaster. Proceedings of the National Academy of Sciences of the United States of America 101, 11685-11688.

Pleil, J.D., Williams, M.A., Sobus, J.R., 2012b. Chemical Safety for Sustainability (CSS): human in vivo biomonitoring data for complementing results from in vitro toxicology--a commentary. Toxicology letters 215, 201-207.

Pope, C.A., 3rd, Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K., Thurston, G.D., 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA : the journal of the American Medical Association 287, 1132-1141.

Prüss-Üstün, A., Corvalán, C., 2006. Preventing disease through healthy environments. Towards an estimate of the environmental burden of disease. World Health Organization, 20 Avenue Appia, 1211 Geneva 27, Switzerland.

Rappaport, S.M., Kupper, L.L., 2008. Quantitative exposure assessment. Stephen Rappaport, El Cerrito, California, U.S.A.

Rappaport, S.M., Smith, M.T., 2010. Epidemiology. Environment and disease risks. Science 330, 460-461.

Rodes, C.E., Lawless, P.A., Thornburg, J.W., Williams, R.W., Croghan, C.W., 2010. DEARS particulate matter relationships for personal, indoor, outdoor, and central site settings for a general population. Atmospheric Environment 44, 1386–1399.

Sexton, K., Adgate, J.L., Church, T.R., Ashley, D.L., Needham, L.L., Ramachandran, G., Fredrickson, A.L., Ryan, A.D., 2005. Children's exposure to volatile organic compounds as determined by longitudinal measurements in blood. Environmental health perspectives 113, 342-349.

Sexton, K., Ryan, A.D., 2012. Using exposure biomarkers in children to compare between-child and within-child variance and calculate correlations among siblings for multiple environmental chemicals. Journal of exposure science & environmental epidemiology 22, 16-23.

Sobus, J.R., McClean, M.D., Herrick, R.F., Waidyanatha, S., Nylander-French, L.A., Kupper, L.L., Rappaport, S.M., 2009a. Comparing urinary biomarkers of airborne and dermal exposure to polycyclic aromatic compounds in asphalt-exposed workers. The Annals of occupational hygiene 53, 561-571.

Sobus, J.R., McClean, M.D., Herrick, R.F., Waidyanatha, S., Onyemauwa, F., Kupper, L.L., Rappaport, S.M., 2009b. Investigation of PAH biomarkers in the urine of workers exposed to hot asphalt. The Annals of occupational hygiene 53, 551-560.

Sobus, J.R., Pleil, J.D., McClean, M.D., Herrick, R.F., Rappaport, S.M., 2010. Biomarker variance component estimation for exposure surrogate selection and toxicokinetic inference. Toxicology letters 199, 247-253.

Sobus, J.R., Tan, Y.M., Pleil, J.D., Sheldon, L.S., 2011. A biomonitoring framework to support exposure and risk assessments. The Science of the total environment 409, 4875-4884.

Steffen, C., Auclerc, M.F., Auvrignon, A., Baruchel, A., Kebaili, K., Lambilliotte, A., Leverger, G., Sommelet, D., Vilmer, E., Hemon, D., Clavel, J., 2004. Acute childhood leukaemia and environmental exposure to potential sources of benzene and other hydrocarbons; a case-control study. Occupational and environmental medicine 61, 773-778.

Tan, Y., Dary, C.C., Chang, D., Ulrich, E.M., Van Emon, J.M., Xue, J., Pleil, J.D., Kenneke, J.F., Sobus, J., Sheldon, L.S., Morgan, M.K., Goldsmith, M., Tornero-Velez, R., Highsmith, R., Fortmann, R.C., Collette, T.W., Zartarian, V.G., 2012a. Biomonitoring - an exposure science tool for exposure and risk assessment. U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-12/039 (NTIS PB2012-112321).

Tan, Y.M., Sobus, J., Chang, D., Tornero-Velez, R., Goldsmith, M., Pleil, J., Dary, C., 2012b. Reconstructing human exposures using biomarkers and other "clues". Journal of toxicology and environmental health Part B, Critical reviews 15, 22-38.

Tornero-Velez, R., Symanski, E., Kromhout, H., Yu, R.C., Rappaport, S.M., 1997. Compliance versus risk in assessing occupational exposures. Risk analysis : an official publication of the Society for Risk Analysis 17, 279-292.

Vogt, R., Bennett, D., Cassady, D., Frost, J., Ritz, B., Hertz-Picciotto, I., 2012. Cancer and non-cancer health effects from food contaminant exposures for children and adults in California: a risk assessment. Environmental health : a global access science source 11, 83.

Williams, R., Case, M., Yeatts, K., Chen, F., Scott, J., Svendsen, E., Devlin, R., 2008. Personal coarse particulate matter exposures in an adult cohort. Atmospheric Environment 42, 6743–6748.

Wolff, M.S., Weston, A., 1997. Breast cancer risk and environmental exposures. Environmental health perspectives 105 Suppl 4, 891-896.

Figure Captions:

Figure 1. Biomonitoring framework integrating exposure to effect continuum and the linkages among measured and modeled estimates. Boxes (lower progression) reflect measurements and triangles (upper progression) reflect exposure and dose estimates based on various models (adapted from Sobus et al. 2011).

Figure 2. Frequency distribution of 220 spot measurements of 1-OH-pyrene in urine from distinct individuals. Red arrow indicates the location of the geometric mean of this particular log-normal distribution.

Figure 3. Frequency distributions for "best" and "worst" case mean values for assessing exceedance past some critical value. If each spot measurement represents that individual's lifetime mean value biomarker, then 10% of the population has a lifetime exceedance past BE = 4400. Based on a population of 220 individuals, if each measurement represents a random spot sample drawn from the global population, then none of the individual means exceed the BE. So, based on a simple data set of spot measures (1 per person), we can only say that at most 10% of the population has lifetime means that exceed the hypothetical level of concern (4400 ng/L). The value could be as low as 0%, but without additional data, we cannot defensibly assign an exceedance value below 10%.

Figure 4. In each case, 25 subjects with 10 measurements/subject were selected at random from data stratified based on ICC outcome. As ICC increases, the within subject scatter decreases and the between subject means distinction increases. For each graph, the overall distribution of spot measures remains the same at GM_g = 1155 and GSD_g = 2.84.

Figure 5. Linear relationships for ICC vs. means distribution parameter, GSD/GM_g for different #'s of measurements per subject (m) using the oh-pyrene data set and hypothetical repeat measures construction. This example is based on empirical measurements of ICC using SAS proc mixed procedure.

Figures 6 a,b. Means distributions for oh-pyrene in urine dependence on different ICC's. a) Total range showing relative position of BE = 4400 ng/L; b) close up of distribution tails showing the areas (to the right of BE) representing the percent of the population expected to have lifetime means greater than BE.