A Proposal for Assessing Study Quality: Biomonitoring, Environmental Epidemiology, and Short-Lived Chemicals (BEES-C) Instrument

Judy S. LaKind,[a] Jon R. Sobus,[b] Michael Goodman,[c] Dana Boyd Barr,[d] Peter Fürst,[e] Richard J. Albertini,[f] Tye E. Arbuckle,[g] Greet Schoeters,[h] Yu-Mei Tan,[b] Justin Teeguarden,[i] Rogelio Tornero-Velez,[b] Clifford P. Weisel[j]

[a]LaKind Associates, LLC; Department of Epidemiology and Public Health, University of Maryland School of Medicine, Department of Pediatrics, Penn State University College of Medicine, Milton S. Hershey Medical Center, 106 Oakdale Avenue, Catonsville, MD 21228 USA lakindassoc@comcast.net
[b]National Exposure Research Laboratory, Human Exposure and Atmospheric Sciences Division, US Environmental Protection Agency, Research Triangle Park, NC 27711 USA Sobus.Jon@epa.gov; Tan.Cecilia@epa.gov; Tornero-Velez.Rogelio@epa.gov
[c]Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Rd, Atlanta, GA 30322 USA mgoodm2@emory.edu
[d]Department of Environmental and Occupational Health, Rollins School of Public Health, Emory University, 1518 Clifton Road, NE, Room 272, Atlanta, GA 30322 dbbarr@emory.edu
[e]Chemical and Veterinary Analytical Institute, Münsterland-Emscher-Lippe (CVUA-MEL) Joseph-König-Straße 40, D-48147, Münster D-48151 Germany Peter.Fuerst@cvua-mel.de
[f]University of Vermont College of Medicine, P.0. Box 168, Underhill Center, VT 05490 USA Ralbert315@aol.com
[g]Population Studies Division, Healthy Environments and Consumer Safety Branch, Health Canada, 50 Colombine Dr., A.L. 0801A, Ottawa ON K1A 0K9 Canada Tye.Arbuckle@hc-sc.gc.ca
[h]Environmental Risk and Health Unit, VITO, Industriezone Vlasmeer 7, 2400 MOL, Belgium; University of Antwerp, Department of Biomedical Sciences, Belgium greet.schoeters@vito.be
[i]Pacific Northwest National Laboratory, 902 Battelle Boulevard, P.O. Box 999, MSIN P7-59 Richland, WA 99352 USA jt@pnnl.gov
[j]Environmental and Occupational Health Sciences Institute, Robert Wood Johnson Medical School, UMDNJ, 170 Frelinghuysen Road, Piscataway, NJ 08854 USA weisel@eohsi.rutgers.edu

Corresponding author: Judy S. LaKind, Ph.D., LaKind Associates, LLC, 106 Oakdale Avenue, Catonsville, MD 21228 USA; PH: +1 410 788 8639; lakindassoc@comcast.net

Abstract

The quality of exposure assessment is a major determinant of the overall quality of any environmental epidemiology study. The use of biomonitoring as a tool for assessing exposure to ubiquitous chemicals with short physiologic half-lives began relatively recently. These chemicals present several challenges, including their presence in analytical laboratories and sampling equipment, difficulty in establishing temporal order in cross-sectional studies, short- and long-term variability in exposures and biomarker concentrations, and a paucity of information on the number of measurements required for proper exposure classification. To date, the scientific community has not developed a set of systematic guidelines for designing, implementing and interpreting studies of short-lived chemicals that use biomonitoring as the exposure metric or for evaluating the quality of this type of research for WOE assessments or for peer review of grants or publications. We describe key issues that affect epidemiology studies using biomonitoring data on short-lived chemicals and propose a systematic instrument – the Biomonitoring, Environmental Epidemiology, and Short-Lived Chemicals (BEES-C) Instrument - for evaluating the quality of research proposals and studies that incorporate biomonitoring data on short-lived chemicals. Quality criteria for three areas considered fundamental to the evaluation of epidemiology studies that include biological measurements of short-lived chemicals are described: 1) biomarker selection and measurement, 2) study design and execution, and 3) general epidemiological study design considerations. We recognize that the development of an evaluative tool such as BEES-C is neither simple nor non-controversial. We hope and anticipate that the instrument will initiate further discussion/debate on this topic.

Key words: BEES-C, biomonitoring, ubiquitous chemicals, short physiologic half-life, evaluation instrument, environmental epidemiology

1. INTRODUCTION

Epidemiological research plays a critical role in assessing the effects of various chemical, physical, biological, radiological, and behavior-related exposures on human health. However, even well-designed and rigorously implemented epidemiological studies that are specifically designed to test causal hypotheses in humans often report conflicting results. Regulatory bodies and consensus panels charged with recommending health policy typically rely on weight-of-evidence (WOE) approaches for evaluating epidemiological research findings. A WOE assessment may be incomplete or misleading if it does not evaluate study quality to ensure that the conclusions are based on the strongest evidence available. In addition, study quality assessments during peer reviews of grant proposals and manuscripts serve to enhance the overall quality of human exposure and health research.

While determination of study quality will always to some extent involve professional judgment, there appears to be an emerging consensus that any evaluation of the strength of epidemiological evidence should rely on agreed-upon criteria that are applied systematically (Vandenbroucke, 2007). These considerations motivated the development and refinement of several study quality assessment tools. Some of these tools (e.g., STROBE [Vandenbroucke et al., 2007]; CONSORT [Moher et al., 2001]) address general issues that apply across disciplines. Other tools were developed specifically for various areas of medicine or life sciences (e.g., STREGA for genetic studies [Little et al., 2009], GRADE for comparative treatment effectiveness research [Owens et al., 2010], and STARD for studies of diagnostic accuracy [Bossuyt et al., 2004]).

In view of the current tendency towards standardization of WOE assessment that incorporates study quality, the relative paucity of instruments for evaluating environmental epidemiology studies – either during development of study design or in review of manuscripts - is notable and difficult to explain. An evaluative scheme focusing on assessing study quality for weight of evidence assessments (Harmonization of Neurodevelopmental Environmental Epidemiology Studies) (Youngstrom et al., 2012) used the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) as the basis for a coding tool (Whiting et al., 2003), but as the name implies, this instrument centered on neurodevelopmental studies. The National Toxicology Program recently developed an approach for assessing study quality (NTP, 2013) and used this to examine the literature on environmental chemicals and diabetes (Kuo et al., 2013); this scheme included assessments of both epidemiologic and toxicology literature and included non-persistent and persistent chemicals but did not incorporate issues specific to biomonitoring of short-lived chemicals.

The lack of a tool that provides systematic guidance on best practices for environmental epidemiological research is an important limitation to regulatory decisions which rely on population-based studies. WOE assessments based on environmental epidemiology data are unique because, unlike other areas of research, experimental studies designed to elicit an adverse outcome in humans are rarely, if ever, ethically possible. Thus, environmental epidemiology studies are almost always observational and are subject to unavoidable uncertainty stemming from various sources. An important source of uncertainty in environmental epidemiology, but also an area of rapid progress, relates to exposure science.

Exposure assessment is a major determinant of the overall data quality in any environmental epidemiology study (Hertz-Picciotto, 1998), including chemicals with short physiologic half lives. Short-lived chemicals are those for which the time required to eliminate one-half of the chemical mass from the body or from a given matrix is on the order of minutes to hours or days. The quality of the exposure assessment for short-lived chemicals is intimately tied to the data's utility in assessing associations with health outcomes as well as to studies using biomonitoring to examine various aspects of exposure. In recent years, exposure science methods have particularly benefited from improvements in the ability to detect environmental chemicals through biomonitoring. Biomonitoring is the measurement of chemicals in various human matrices such as blood, urine, breath, milk and hair. Biomonitoring data integrate exposure from all routes (oral, inhalation, dermal, trans-placental) and are valuable for: (1) establishing population reference ranges; (2) identifying unusual exposures for subpopulations; (3) evaluating temporal variability and trends within a population; (4) validating questions designed to estimate individual exposure; and (5) examining associations with health outcomes in epidemiologic studies.

Epidemiologic research with biomonitoring as the basis for measuring exposure for persistent organic pollutants and metals has been conducted for decades. By contrast, biomonitoring of ubiquitous chemicals with short physiologic half-lives (e.g., benzene, phthalates, certain pesticides) began relatively recently, and these chemicals present several new challenges as interpretation of data on these chemicals is complicated by variability in exposure and the ubiquitous nature of many of these chemicals, including in analytical laboratories and sampling equipment. These chemicals also present challenges when selecting the matrix to be used in the research. To date, the scientific community has not developed a set of systematic guidelines for implementing and interpreting biomonitoring studies of these chemicals. Similarly, there is no published method for evaluating the quality of this type of research for WOE assessments or for peer review of grants or publications.

This knowledge gap was the specific focus of the 2013 international workshop "Best Practices for Obtaining, Interpreting and Using Human Biomonitoring Data in Epidemiology and Risk Assessment: Chemicals with Short Biological Half-Lives." The workshop brought together an expert panel from government, academia, and private institutions specializing in analytical chemistry, exposure and risk assessment, epidemiology, medicine, physiologically-based pharmacokinetic (PBPK) modeling, and clinical biomarkers. The aims of the workshop were to (i) describe the key issues that affect epidemiology studies using biomonitoring data on chemicals with short physiologic half lives, and (ii) develop a systematic scheme for evaluating the quality of research proposals and studies that incorporate biomonitoring data on short-lived chemicals.

Quality criteria for three areas considered to be fundamental to the evaluation of epidemiology studies that include biological measurements of short-lived chemicals are described in this paper: 1) biomarker selection and measurement, 2) study design and execution, and 3) general epidemiological study design considerations. Key aspects of these topic areas are discussed and are then incorporated into a proposed evaluative instrument – the Biomonitoring, Environmental Epidemiology, and Short-Lived Chemicals (BEES-C) instrument - organized as a tiered matrix (Table 1). Some aspects of the proposed evaluative instrument include study design elements

that are relevant to epidemiology studies of both persistent and short-lived chemicals. In fact, aspects of widely accepted instruments such as STROBE have intentionally been weaved into the evaluative instrument proposed here (Little et al., 2009; Vanderbrouchke et al., 2007; Gallo et al., 2011). (STROBE offers guidance regarding methods for improving on reporting of observational studies and for critically evaluating these studies; STROBE is designed to be used by reviewers, journal editors and readers [(Vandenbroucke et al., 2007)].) While both established and novel aspects of this instrument are critical to assessing the quality of a study using biomonitoring of short-lived chemicals as an exposure assessment approach, the primary objective of this communication is to cover critical aspects of studies of short-lived chemicals; these are described more fully in the text.

The list of quality issues that could be used to evaluate a given study is long; a tension exists between the development of an all-inclusive but unwieldy instrument versus a more discriminating and utilitarian instrument that includes only the most important issues (focusing on those research aspects that are unique – or of particular importance - to short-lived chemicals). We opted for the latter in developing the proposed BEES-C Instrument. The instrument can be applied to studies that examine the relation between exposure and health outcome as well as to studies using biomonitoring data to various aspects of exposure (e.g., temporal and spatial trends). The issues raised here and addressed by the BEES-C instrument cut across multiple disciplines that involve biological measurements of short-lived chemicals, including occupational studies and nutritional epidemiology.

The features of short-lived chemicals in environmental epidemiology studies that require special attention are: the number and timing of samples taken in order to represent the relevant exposure window for the health outcome of interest; the ubiquitous use of many of these chemicals in currently manufactured products, including personal care products, laboratory equipment, dust, food, etc., which introduces special needs for avoidance of sample contamination; choice of appropriate biological matrix; and the ability to measure a large number of chemicals in one sample, increasing the need for attention to full reporting and issues related to multiple comparisons. These are discussed more fully in the following sections, with examples given for each issue. While most of the instrument topics pertain to biomarkers of exposure, biomarkers of effect are described when relevant.

2. USING THE BEES-C INSTRUMENT

The BEES-C instrument can serve multiple purposes including: aiding researchers in the development of study design, reviewing grant proposals, peer reviewing manuscripts, and conducting WOE assessments.

2.1 Intended uses of BEES-C

The ultimate goal of the BEES-C tool is to assist researchers in improving the overall body of literature on studies of short-lived chemicals in humans. The BEES-C instrument is not intended to be used: (i) to discourage researchers from conducting hypothesis-generating research, or (ii) to preclude lower-tiered studies from being included in WOE assessments.

As with any type of evaluative instrument, professional judgment must be part of the evaluative process, both in terms of tiering and for determining which aspects of the instrument are relevant to a given study.

In the sections below, we describe the key aspects of BEES-C along with examples. Here we discuss recommendations for utilizing BEES-C. While the preponderance of the topics covered by this instrument would pertain to human biomonitoring studies that are part of epidemiological research on associations between biomarkers of exposure and some measure of effect (e.g., biomarker of effect, physician-diagnosed disease), only a portion of the BEES-C instrument will be applicable to human biomonitoring studies designed for other purposes (e.g., exposure assessment for temporal or spatial trend analyis).

2.2 How to use BEES-C

Table 1 is organized according to aspects of study design (rows) and evaluative tiers (columns). For each study under review, critical aspects are assessed row by row and the appropriate cell is color-coded (Figure 1), with Tier 1 indicating the highest quality. This allows the researcher/reviewer to obtain an overall picture of study quality. The user of this instrument should provide justification for each decision made (Table 1); this will enhance transparency in the process. The BEES-C instrument can be used: (i) as an instrument by researchers evaluating their proposed study design to ensure that the study quality is maximized; (ii) by reviewers of manuscripts and publications to systematically assess the quality of the research and identifying areas where quality could be improved; (iii) by those performing systematic reviews for evaluating study quality in order to inform decision-making (e.g., Is a study of sufficiently high quality to use in developing regulatory standards? Should a study be included in a meta-analysis?); and (iv) by others wishing to incorporate BEES-C into their currently existing review schemes. For example, many of the issues in our proposed approach that are specifically applicable to short-lived chemicals are not yet part of the draft Office of Health Assessment and Translation Approach (NTP, 2013) but could be incorporated into their approach for conducting “literature-based evaluations to assess the evidence that environmental chemicals, physical substances, or mixtures (collectively referred to as "substances") cause adverse health effects.”

Table 1: Biomonitoring, Environmental Epidemiology, and Short-Lived Chemicals (BEES-C) Instrument: Evaluative instrument for assessing quality of epidemiology studies involving biomonitoring of chemicals with short physiologic half-lives. Evaluative criteria cover several aspects of environmental epidemiology research with biomonitoring as the exposure metric (acronyms defined at bottom of table). The justification column is used to increase transparency in the process of decision-making.

STUDY ASSESSMENT COMPONENTS	TIER 1	TIER 2	TIER 3	Justification
Biomarker Selection and Measurement				
Biological relevance (Parent/surrogate relationship) Exposure biomarker	Biomarker in a specified matrix has accurate and precise quantitative relationship with external exposure, internal dose, or target dose.	Evidence exists for a relationship between biomarker in a specified matrix and external exposure, internal dose, or target dose.	Biomarker in a specified matrix is a poor surrogate (low accuracy and precision) for exposure/dose.	
Effect biomarker	Bioindicator of a key event in an AOP.	Biomarkers of effect shown to have a relationship to health outcomes but the mechanism of action is not understood.	Biomarker has undetermined consequences (e.g., biomarker is not specific to a health outcome).	
Specificity	Biomarker is derived from exposure to one parent chemical.	Biomarker is derived from multiple parent chemicals with similar adverse endpoints.	Biomarker is derived from multiple parent chemicals with varying types of adverse endpoints.	
Method sensitivity (detection limits)	Limits of detection are low enough to detect chemicals in a sufficient percentage of the samples to address the research question.	NA	Frequency of detection too low to address the research hypothesis.	
Biomarker stability	Samples with a known	Samples have known	Samples with either unknown	

	history and documented stability data or those using real-time measurements.	losses during storage but the difference between low and high exposures can be qualitatively assessed.	history and/or no stability data for analytes of interest.	
Sample contamination	Samples are contamination-free from time of collection to time of measurement (e.g., by use of certified analyte-free collection supplies and reference materials, and appropriate use of blanks both in the field and lab). Research includes documentation of the steps taken to provide the necessary assurance that the study data are reliable.	Study not using/documenting these procedures.	There are known contamination issues and no documentation that the issues were addressed.	
Method requirements	Instrumentation that provides unambiguous identification and quantitation of the biomarker at the required sensitivity (e.g., GC-HRMS, GC-MS/MS, LC-MS/MS).	Instrumentation that allows for identification of the biomarker with a high degree of confidence and the required sensitivity (e.g., GC-MS, GC-ECD).	Instrumentation that only allows for possible quantification of the biomarker but the method has known interferants (e.g., GC-FID, spectroscopy).	
Matrix adjustment	Study includes results for adjusted and non-adjusted concentrations if adjustment is needed.	Study only provides results using one method (matrix-adjusted or not).	No established method for adjustment (e.g., adjustment for hair)	
Study Design and Execution				
Temporality	Established time order between exposure and outcomes; relevant interval between the exposure and the outcome or	Established time order between exposure and outcome, but no consideration of relevant exposure windows.	Study without an established time order between exposure and outcome.	

	reconstructed exposure and appropriate consideration of relevant exposure windows.			
Exposure variability and misclassification	Sufficient number of samples. Error considered by calculating measures of accuracy (e.g., sensitivity and specificity) and reliability (e.g., ICC). If one sample is used, there is evidence that errors from a single measure are negligible.	More than one sample collected, but without explicit evaluation of error.	Exposure based on a single sample without considering error.	
General Epidemiological Study Design Considerations				
Study rationale	Studies designed specifically to evaluate an *a priori* formulated hypothesis.	Studies using existing samples or data to evaluate an *a priori* formulated hypothesis.	Data mining studies without a pre-specified hypothesis; multiple simultaneous hypothesis testing.	
Study participants	Population-based unbiased selection protocol; high response rate and/or low loss to follow-up.	Population-based unbiased selection protocol; low response rate and/or high loss to follow-up.	Methods of sample selection, and response/loss to follow-up rates are not reported.	
Data analysis	Clear distinction between causal and predictive models; adequate consideration given to extraneous factors with assessment of effect modification and adjustment for confounders; sensitivity analyses.	Adequate consideration of extraneous factors, but without sensitivity analyses.	Inadequate control for extraneous factors.	
Reporting	Study clearly states its aims and allows the reader to evaluate the number of tested hypotheses (not just	Conclusions appear warranted, but the number of tested hypotheses is unclear (either not	Studies that selectively report data summaries and lack transparency in terms of methods or selection of	

	the number of hypotheses for which a result is given). If multiple simultaneous hypothesis testing is involved, its impact is assessed, preferably by estimating PFP or FP:FN ratio. There is no evidence of outcome reporting bias, and conclusions do not reach beyond the observed results.	explicitly stated or difficult to discern) and/or there is no consideration of multiple testing.	presented results.	

AOP = adverse outcome pathways; FP = false positive; FN = false negative; GC-HRMS = gas chromatography/high-resolution mass spectrometry; GC-MS = gas chromatography/mass spectrometry; GC-ECD = gas chromatography-electron capture detector; GC-FID = gas chromatography-flame ionization detector], ICC = intra-class correlation coefficient; NA = not applicable; PFP = probability of false positive

Figure 1. Example of quality comparison of two hypothetical studies with biomonitored short-lived chemicals using the BEES-C instrument. For each hypothetical study under review, critical aspects are assessed row by row and the appropriate cell is color-coded, allowing the researcher/reviewer to obtain an overall picture of study quality. Text in cells has been removed for readability.

Hypothetical Study 1

STUDY ASSESSMENT COMPONENTS	TIER 1	TIER 2	TIER 3
Biomarker Selection and Measurement			
Biological relevance Exposure biomarker			
Effect biomarker			
Specificity			
Method sensitivity			
Biomarker stability			
Sample contamination			
Method requirements			
Matrix adjustment			
Study Design and Implementation			
Temporality			
Exposure variability and misclassification			
General Epidemiological Study Design Considerations			
Study rationale			
Study participants			
Reporting			
Data analysis			

Hypothetical Study 2

STUDY ASSESSMENT COMPONENTS	TIER 1	TIER 2	TIER 3
Biomarker Selection and Measurement			
Biological relevance Exposure biomarker			
Effect biomarker			
Specificity			
Method sensitivity			
Biomarker stability			
Sample contamination			
Method requirements			
Matrix adjustment			
Study Design and Implementation			
Temporality			
Exposure variability and misclassification			
General Epidemiological Study Design Considerations			
Study rationale			
Study participants			
Reporting			
Data analysis			

Implicit in this study quality evaluative instrument is that the manuscript or proposal will explicitly report on each of the issues below. In other words, in order to assess whether the study meets the criteria for a given tier, the information on that issue must be clearly described. For studies relying on previously-published biomonitoring data (e.g., US National Health and Nutrition Examination Survey [NHANES]), the same reporting requirements must be met. Authors should be explicit in their description of methods, including pertinent details such as limit of detection for the study, relative standard deviation and relevant quality control parameters.

The lack of numeric scoring for this process is intentional. There will no doubt be instances where a study is of high quality for most components, but has not addressed a key issue that substantially reduces confidence in the study results. An overall high "score" would mask this problem. Instead, we propose a qualitative approach that increases flexibility.

A final note: We are unaware of studies that would be categorized as Tier 1 for all aspects of the evaluation. While a study that falls into Tier 1 for all aspects is certainly a goal and would provide robust data, it is the case that most studies will contain aspects that would be considered Tier 2 or 3. Depending on the users' intent for the study data, this may not be problematic for certain evaluative issues. On the other hand, there are some issues for which a Tier 3 designation would render the study of low utility (e.g., inability to demonstrate samples were free of contamination).

3. COMPONENTS OF BEES-C

We first describe BEES-C components specifically related to short-lived biomarkers. This is followed by aspects of BEES-C that pertain to more general epidemiological study design issues.

3.1 Biomarker Selection and Measurement

A biomarker/biological marker has been defined as an "indicator of changes or events in biological systems. Biological markers of exposure refer to cellular, biochemical, analytical, or molecular measures that are obtained from biological media such as tissues, cells, or fluids and are indicative of exposure to an agent" (Zartarian et al., 2005). Thus, biomarkers can be used to assess exposure to a chemical by measuring the amount of that chemical or its metabolite in the body. In addition, biomarkers can be used as indicators of health effects. Many biomarkers of exposure and effect are short-lived, and both types of biomarkers are commonly used in human research on exposure to – and health effects from – environmental chemicals. While this evaluative tool is predominantly focused on biomarkers of exposure, many of the principles elucidated here also apply to biomarkers of effect.

As a general rule, studies designed to observe associations between exposure and health effects are more defensible if appropriate and well-established biomarkers are used as exposure and/or health endpoint surrogates. There is general consensus on certain criteria that should be met for biomarkers to be considered high-quality (NRC, 2006; Zelenka et al., 2011). Some of these criteria are based on the inherent qualities of the biomarkers (e.g., its relevance to chemical exposure and/or biological relevance). Other criteria pertain to the measurement of the biomarker – that is, the accuracy and precision of methods used to quantify the biomarker, the stability of the biomarker during storage, the possibility for sample contamination leading to errors in biomarker quantitation, and the need to adjust for biological matrix effects that might introduce measurement error. Critical aspects of biomarker selection and measurement are described in the following subsections and the proposed tiering scheme for BEES-C is shown in Table 1.

3.1.1 Relevance

Source-to-outcome continuums are frequently used to demonstrate the path of a chemical from generation, to human contact, to target dose and subsequent molecular, cellular, organ, organism, and population response. Biomarkers are sometimes used as a means to empirically characterize exposure, dose, and biological response. In this section we consider both biomarkers of exposure (i.e., a parent chemical, metabolite, or interaction product at a target [WHO, 2001]) and biomarkers of effect (i.e., a measureable biochemical or physiological alteration that is associated with a health outcome [WHO, 2001]) as important components of epidemiological studies of associations between exposure and health outcome.

Biomarkers of exposure: Epidemiologic research can be hypothesis-driven or more geared towards hypothesis-generation. In the latter case, the most suitable biomarker of exposure is one that is an accurate and precise surrogate of external exposure or internal dose. When a strong biological rationale exists, and a biological "target" is known, the most suitable biomarker is one that is directly measured at the target (molecular, cellular, or organ level), or is an accurate and precise surrogate of target dose.

Ideally, a clear understanding of the quantitative linkages between exposure, dose, and biomarker levels will exist for any biomarker that is used in an epidemiological study. Considering the invasive nature of target tissue sampling, most biomarker-based epidemiological studies utilize samples of blood, urine, hair, or other easily-accessible matrices. Elucidating quantitative relationships between biomarker measurements from these matrices and exposure/dose levels requires an understanding of chemical absorption, distribution, metabolism, and elimination (ADME); these processes are frequently described using pharmacokinetic (PK) models, or physiologically-based pharmacokinetic (PBPK) models. Prior to the use of biomarkers in an epidemiological study, a solid understanding of chemical ADME should exist, as well as the intrinsic (e.g., genetics, life-stage, pregnancy, gender) and extrinsic (e.g., diet, medication, medical conditions) factors that are likely to affect ADME. Furthermore, for short-lived biomarkers, it is important to know specific timing details (e.g., time of day, time since last meal for those chemicals associated with dietary exposure, time since last urine void) in relation to sample collection. Ideally, the relationships between biomarker concentration and exposure/dose levels, and the effects of intrinsic, extrinsic, and timing factors on these relationships, will be thoroughly evaluated before the biomarker is used in an epidemiological study. Critical information that is needed to properly interpret the biomarker (with respect to exposure/dose) should then be collected and carefully evaluated as part of the study. The costs and benefits of each biomarker of exposure should be carefully examined and interpreted as part of any epidemiological evaluation.

It is important to note that matrix selection is an integral component of exposure and/or epidemiology research, and multiple factors must be considered including measurement capability, contamination issues, and target analyte association with exposure or health outcome. BEES-C addresses each of these issues separately.

Short-lived chemical example : Bisphenol A (BPA) is measured in urine in the free form (parent), as sulfate- or glucoronide-bound conjugates, or as a combination (total BPA) of the free and conjugated forms (Harthé et al. 2012; LaKind et al., 2012a; Völkel et al. 2008; Ye et al. 2005). Several recent studies have examined endocrine-related health outcomes associated with

BPA exposure. The most biologically-relevant biomarker is the free (parent) BPA, because only parent BPA is considered active in terms of estrogenicity (EPA, 2013; WHO, 2011). The quantification of free BPA in urine is analytically challenging, however, as only a small fraction of BPA is present in the non-conjugated form (Ye et al., 2005). Given this limitation, measurements of conjugated or total BPA may be useful surrogates of free BPA. Specifically, if there is small variation in the ratio of free to conjugated BPA within and between individuals (with respect to the variation in exposure levels), then conjugated or total BPA may be an accurate and precise surrogate of free BPA, and of BPA exposure in general. This example underscores the importance of understanding relationships between exposure and biomarkers, different types of biomarkers (parent vs. metabolites in their respective matrices), and biomarkers and biological targets, while ensuring that the appropriate research question is addressed. It further highlights the possibility of trade-offs when selecting an individual biomarker of exposure (for BPA, biological relevance could be optimized at the expense of ability to detect the chemical).

Study evaluation (Table 1): A Tier 1 biomarker of exposure in a specified matrix is an accurate and precise surrogate of target dose (for hypothesis-driven studies with a known target) or of external exposure (for studies without a known target). For a Tier 2 biomarker, evidence exists for a relationship between the biomarker in a specified matrix and external exposure, internal dose, or target dose. A Tier 3 biomarker in a specified matrix is a poor surrogate (low accuracy and precision) for exposure/dose.

Biomarkers of effect: It can be challenging in epidemiological studies to perform meaningful comparisons of short-lived biomarker measurements and long-term health outcomes. Particularly in cross-sectional studies, a key assumption is that current biomarker levels reflect past exposures during time windows that were relevant for disease onset. Biomarkers of effect offer a means to evaluate exposure-response relationships in target populations, during critical time windows, prior to disease onset. Findings are interpreted based on the strength of association between biomarkers of exposure and effect, and between biomarkers of effect and the adverse health outcome.

The progression from an exposure event to an adverse health effect can be defined using adverse outcome pathways (AOPs) (Ankley et al., 2010). The AOP for a particular health outcome begins with a molecular initiating event at a target within the body. Effects at the molecular target, initiated by exposure events, progress to effects at the cellular, tissue, and organ levels, and ultimately to the whole organism. "Key events" are intermediate steps along the AOP that can be experimentally monitored to evaluate progression along the AOP. Measurements of these key events in accessible biological media from living intact organisms are called bioindicators. Bioindicators are considered ideal biomarkers of effect because they reflect a biological function linked to a specific adverse outcome; they "provide a high degree of confidence in predicting the potential for adverse effects in an individual or population" (www.epa.gov/pesticides/science/biomarker.html). Biomarkers of effect categorized as "Undetermined Consequences" reflect a less certain pathway linking alterations to any specific disease outcome (www.epa.gov/pesticides/science/biomarker.html). Predictions of outcomes therefore, for either individuals or populations, are less certain when using these biomarkers in place of bioindicators.

Study evaluation (Table 1): A Tier 1 biomarker of effect is a bioindicator of a key event in an AOP. A Tier 2 biomarker of effect has been shown to have a relationship to health outcomes but the mechanism of action is not understood. Biomarkers of effect that have undetermined consequences are considered Tier 3.

3.1.2 Specificity

A single biomarker of exposure may be derived from multiple parent chemicals, making assessments of exposure to the parent chemical difficult to ascertain (Barr et al., 1999, 2006; Barr and Needham 2002). In terms of exposure assessment and interpretation of epidemiological research, this is especially problematic if the parent chemicals have different toxicities or modes of action. Further, an example of interference with assessing exposure to a parent chemical is the situation in which one of the metabolites also can be found in the environment (an exogenous source).

Short-lived chemical example: 3-phenoxybenzoic acid (3PBA) is an example of a short-lived chemical that highlights the importance of evaluation of specificity when assessing study quality. 3PBA is a metabolite of at least 18 synthetic pyrethroids (Barr et al., 2010; Leng et al., 1997) and is also a potential metabolite of the 3PBA environmental degradate 3-phenoxybenzyl alcohol. Thus, urinary 3PBA measurements represent exposure to multiple insecticides with varying degrees of neurotoxicity, in addition to exposure to an environmental degradate that is not known to be neurotoxic (Barr et al., 2010). Urinary 3PBA measurements can therefore provide a conservative estimate of pyrethroid exposure; however, it likely would not provide an accurate exposure estimate for neurotoxic effects related to pyrethroid insecticide exposure in the absence of additional exposure data. Thus, finding a relation between neurotoxicity and exposure would be more difficult since the true exposures are unknown.

Study evaluation (Table 1): A Tier 1 study includes a biomarker of exposure that is derived from exposure to one parent chemical. A Tier 2 study includes a biomarker derived from multiple parent chemicals with similar types of adverse endpoints. A Tier 3 study includes a biomarker derived from multiple parent chemicals with varying types of adverse endpoints.

3.1.3 Method sensitivity

The biomarker should be appreciably present in the matrix being analyzed (Calafat and Needham, 2008). A biomarker that is frequently non-detectable in a matrix - irrespective of exposure - is undesirable in environmental epidemiologic research as the results may be of limited utility.

Short-lived chemical example: Several polycylic aromatic hydrocarbons (PAHs) with four or more rings are suspected or known human carcinogens (e.g., benzo[a]pyrene). Standard analytical methods (e.g., GC-MS [gas chromatography/mass spectrometry] or LC-MS/MS [liquid chromatography-tandem mass spectrometry]) are often not sufficiently sensitive for quantifying metabolites of these PAHs in accessible media (e.g., urine) (Bouchard and Viau, 1997), thus hindering epidemiological investigations. Biomarkers of smaller PAHs, including naphthalene, phenanthrene and pyrene, have been evaluated as surrogates of the larger carcinogenic species (Bouchard et al., 1998, Viau et al., 1999; Sobus et al., 2009; Withey at al.,

1991). These surrogates offer a means to overcome analytical limitations, but must be thoroughly evaluated for their ability to reflect exposure to the target species, to gauge co-occurrence among the PAHs, and to evaluate information on correlates of exposure sources.

Study evaluation (Table 1): A Tier 1 study method has limits of detection low enough to detect chemicals in a sufficient percentage of the samples to address the research question (e.g., 50-60% detectable values if the research hypothesis requires estimates of both central tendencies and upper tails of the population concentrations) (Barr et al., 2010; Zota et al., 2014). There is no Tier 2 for this component. A Tier 3 study has too low a frequency of detection to address the research hypothesis.

3.1.4 Biomarker stability

The biomarker should be stable in a given matrix over the time of storage and use (Barr et al., 2005a). Stability of the sample should be documented. Studies using samples that have undergone freeze/thaw cycles should demonstrate the stability of those samples. Time from collection of sample to measurement should be documented.

Short-lived chemical example: While persistent organic pollutants are usually stable in blood products stored indefinitely if frozen at -20ºC or below, non-persistent chemicals may be less stable in blood. For example, current-use pesticides are highly reactive and can easily degrade in blood enzymatically (Barr et al., 1999). Blood preserved with EDTA minimizes esterase activity but the measurement should be made within a few months after collection. Thaw/refreeze cycles or thawing samples in hot water can also cause degradation. The use of long-archived urine or blood samples may provide data on historically collected samples (e.g., NHANES III samples) but many have experienced thaw/refreeze cycles that can result in degradation of sensitive chemicals or contamination of the sample itself. Small, multiple aliquots of a single sample should be stored to be able to confirm the stability of historic samples. Losses of biomarkers can also occur from binding to the walls of the containers and from volatilization. While plastic containers are inexpensive and easy to handle and freeze compared to glass, they can be a source of contamination of some chemicals. In addition, they can absorb both metals and organic compounds resulting in underestimation of chemical concentration. Storage studies using spiked matrices at levels consistent with those expected to be found in the actual sample or the addition of stable isotopically labeled compounds to samples prior to storage should be done to validate that there are no losses during storage or in thaw-refreeze cycles.

Study evaluation (Table 1): A Tier 1 study would include samples with a known history and documented stability data. Tier 2 studies have known losses during storage but the difference between low and high exposures can be qualitatively assessed (i.e., for the purposes of the study, it is sufficient to bin study participants as having either low or high exposure). Tier 3 studies use samples with either unknown history and/or no stability data for the analyte(s) of interest.

3.1.5 Sample contamination

This BEES-C evaluative criterion is one of the most critical criteria for evaluating studies measuring ubiquitous short-lived chemicals. This is because the likelihood of sample contamination from the time of collection to the time of measurement has been demonstrated for many of these chemicals, this in spite of great lengths taken to avoid contamination. A wide

range of chemicals with short physiologic half lives are not only environmentally ubiquitous but may also be present in the sampling and analytical equipment used in epidemiological research. Thus, extreme care is necessary in order to avoid/prevent sample contamination during all phases of a study from sample collection to sample measurement (Barr et al., 1999; Calafat and Needham, 2008, 2009; Needham et al., 2007). During sample collection, supplies containing the target chemical or exposing the collection materials or matrix to environmental media (e.g., air or water) can falsely elevate the measured concentrations. Even with precautions, studies have reported difficulties with analytic contamination, contributing to uncertainty in interpretation of study results.

Short-lived chemical example : Ye et al. (2013) note that despite their best efforts, samples at the Centers for Disease Control Prevention laboratory were contaminated with triclosan; the source of the contamination was ultimately identified as a triclosan-containing handsoap used by a technician. Similarly, several research groups have noted the difficulties in attempting to measure BPA in blood samples, in part, because of contamination (including in solvents and reagents) despite great care taken to avoid such contamination (Calafat et al., 2013; Markham et al., 2010; Teeguarden et al., 2011; Ye et al., 2013).

Study evaluation (Table 1): A Tier 1 study ensures the samples are contamination-free from time of collection to time of measurement (e.g., by use of certified analyte-free collection supplies and reference materials, and appropriate use of blanks both in the field and lab). The research will include documentation of the steps taken to provide the necessary assurance that the study data are reliable and accurate. Any study not using/documenting these procedures is categorized as Tier 2. In a Tier 3 study, there are known contamination issues and no documentation that the issues were addressed.

3.1.6 Method requirements

The quality of a biomarker for assessing exposure is largely dependent upon the quality of the method used for measurement. This can be a difficult aspect of biomarker measurement to evaluate. For example, a laboratory's participation and success in a proficiency testing exercise may seem to be a reasonable test for a Tier 1 study; however, many proficiency testing studies have tolerance ranges that can vary by 200% (i.e., an "acceptable" analyte concentration value can be +/- 200% of the true value). In general, the study methods should have appropriate instrumentation and describe the accompanying procedures (e.g., QC, method robustness, presence of confirmation ions, use of isotope dilution).

Study evaluation (Table 1): A Tier 1 study includes instrumentation that provides unambiguous identification and quantitation of the biomarker at the required sensitivity (e.g., GC-HRMS [gas chromatography/high-resolution mass spectrometry], GC-MS/MS, LC-MS/MS). A Tier 2 study uses instrumentation that allows for indentification of the biomarker with a high degree of confidence and the required sensitivity (e.g., GC-MS, GC-ECD [gas chromatography-electron capture detector]). A Tier 3 study uses instrumentation that only allows for possible quantification of the biomarker but the method has known interferants (e.g., GC-FID [gas chromatography-flame ionization detector], spectroscopy).

3.1.7 Matrix adjustment

Biomarkers are most commonly measured and reported in units of concentration; that is, mass of biomarker/volume of biological media. There are strong effects of variable urine output (driven by diet, exercise, hydration, age, disease state, etc.) on urinary biomarker concentration, and of blood volume and fat content on blood biomarker concentration. Urine biomarker concentrations have been normalized across and within subjects to correct for variable urine dilution using creatinine concentration (derived from creatine phosphate breakdown in muscle), specific gravity, urine output, and other methods, though uncorrected urinary levels in spot samples without auxiliary information are commonly reported and utilized in assessments of exposure and relationship to health outcomes (Barr et al., 2005b; LaKind and Naiman, 2008, 2011; Lorber et al., 2011; Meeker et al., 2005). There is no current consensus on the best method(s) for "correcting" urinary biomarkers measurements for variable urine dilution. Minimally, both the volume-based and a corrected (creatinine and/or other method) concentrations should be provided to allow appropriate comparison across studies. It is also instructive to obtain a full volume void and elapsed time between voids.

Blood-based biomarker levels have been reported in whole blood, serum, plasma and as lipid-adjusted values. The method used to determine the lipid correction or to separate the different components of the blood fluid should be provided and all concentrations, when available, should be reported (e.g., whole volume and lipid-adjusted). Similarly, issues related to fasting samples and serum lipid adjustment in measures of lipophilic chemicals must be considered (Schisterman et al., 2005). The validity of lipid and other tissue component adjustments have not been established for certain short-lived chemicals such as current use pesticides. In these instances, the whole-volume concentrations and adjusted concentrations should be reported with a notation that adjustment validity has not been established. In addition, plasma volume increases in pregnancy (and may also increase for some pre-existing diseases or underlying health conditions) and may also need to be considered when comparing plasma concentrations across pregnancy or populations (Hytten, 1985).

Information about the sample collection requirements and matrix treatment is important when comparing data across studies or to reference ranges. Studies by different governmental agencies (e.g., the European Union, specific European countries, US NHANES, Canadian Health Measures Survey, Consortium to Perform Human Biomonitoring on a European Scale, state-based HANES) and other large biomonitoring data repositories may have different protocols for collecting and processing samples that can alter the matrix and reported biomarker concentrations. For example, instructions given to the participant about fasting prior to sample collection can minimize the lipid content in blood thus minimizing a lipophilic biomarker concentration in a sample (Barr et al., 2005a), and these instructions are not necessarily the same from country to country (LaKind et al., 2012a). Similarly, collection of a first morning urine void may be more concentrated in matrix components than a simple spot sample which may alter our ability to detect or differentiate an analyte (Kissel et al., 2005; Scher et al., 2007). Further, first morning void collection can result in a bias (systematic error) in the data due to the relationship between previous exposure and sample collection and measurement; this is especially important for chemicals for which diet is a predominant route of exposure as the void would be collected after overnight fasting. Blood plasma collected with EDTA versus heparin as an anticoagulant may alter the properties of the matrix (Barr et al., 2005a). Differences in collection requirements and sample processing (as well as health conditions of study participants - such as

kidney disease - that could affect biomarker concentrations) need to be reported, considered and weighed accordingly when results are compared across studies.

Study evaluation (Table 1): We recognize that the best practice for matrix adjustment is intimately associated with the hypothesis to be tested and the specific chemical of interest, and that consensus in this area has not yet been reached. However, adjustment can have a significant effect on study outcome. We therefore propose that a Tier 1 study would provide results for adjusted and non-adjusted concentrations (if adjustment is needed), thereby allowing the reader to reach their own conclusions about the impact of matrix adjustment. A Tier 2 study is one that only presents the results using one method (matrix-adjusted or not). A Tier 3 study includes measurements of a chemical in a matrix that does not yet have a validated adjustment method.

3.2 Study Design and Execution

Considerations of both study design and exposure variability and misclassification are especially important for short-lived chemicals.

3.2.1 Epidemiology study design

Studies that explore associations between biomonitoring data on short-lived chemicals and disease present a unique set of challenges because blood or urine levels of biomarkers typically reflect recent exposures that occurred just hours or at most days ago, and the timing of the exposure relative to the biomarker sample collection is usually not known. Yet most health outcomes of interest are chronic conditions (e.g., obesity, hypertension, or measures of reproductive function) that may require years to decades to develop. For this reason, evaluation of causal hypotheses in studies that measure short-lived chemicals is complicated, and in some circumstances, may not be feasible. A critical and, perhaps the only inarguable, property of a causal association is temporality, meaning that a claim of causation must be supported by an observation of the putative causal exposure preceding the outcome (Potischman and Weed, 1999; Rothman and Greenland, 2005; Weed and Gorelic, 1996; Weed, 1997).

Establishing temporality is only possible in “incidence” studies, which identify health-related events such as new cases of disease at the time of onset or a change in a health-related measure compared to baseline (Pearce, 2012). Incidence studies may be experimental (e.g., clinical trials) or observational (cohort or case-control with ascertainment of incident cases). Regardless of design, however, the main feature of incidence studies is the ability to establish the time of disease onset (or at least the time of diagnosis), which may then allow for an assessment of the sequence of exposure and outcome. In a situation when exposure levels may rapidly change over time, a useful approach is a longitudinal study that assesses the relation between repeated measures of exposure and repeated measures of health biomarkers.

Although the ability to establish the temporal relation is critical for assessing causation, a separate study design issue in environmental epidemiology research is the interval between the exposure and the outcome under study. In order to use human biomonitoring data in etiologic research, exposures should be measured at times which are relevant for disease onset. While this is not a simple task, there are examples of successful biomonitoring studies that have examined exposures of persistent chemicals during relevant time windows and correlated those exposures

with development of specific adverse outcomes. For example, blood lead levels reflect exposures during the preceding 5-6 weeks; and well-conducted epidemiological studies have been able to link the blood levels in children to adverse effects on cognitive capacity (Lanphear et al. 2000). For chemicals with short half-lives, however, the interval between the relevant exposure and disease development is often difficult to assess. Study design – along with exposure misclassification discussed later in this paper – are the most critical and underexplored aspects of biomonitoring studies of short-lived chemicals.

Establishing temporality is much more difficult in a "prevalence" study compared to an "incidence" study, which makes it challenging to draw conclusions about causal associations. A typical prevalence study relies on cross-sectional design, which ascertains the exposure and disease information simultaneously (Rothman and Greenland, 1998). When research is focused on short-lived chemicals, many case-control studies - even if they use incident cases - are difficult to interpret because the biomarker levels reflect recent exposures that typically follow rather than precede disease onset. The notable exception is a study that uses samples collected and stored for future use, as is done in nested case-control or case-cohort studies (Gordis, 2008).

Short-lived chemical example: In a recent review of the epidemiology literature on phthalate metabolites (Goodman et al., 2014) and their association with obesity, diabetes, and cardiovascular disease, most of the studies were cross-sectional in design. The study results were inconsistent across outcomes and lack of temporality was identified as a key limiting factor in the ability to discern relationships between prior exposures to phthalate metabolites and consequent health outcomes.

Study evaluation (Table 1): Tier 1 studies are incidence studies that involve a follow-up time period or a longitudinal analysis of repeated measures and allow the establishment of both the time order and the relevant interval between the exposure and the outcome (Table 1). A Tier 2 study would include incidence studies in which exposure preceded the outcome, but the specific relevant windows of exposure are not considered. The least informative (Tier 3) studies are those that examine the association between current exposure (e.g., blood level of a chemical) and frequently measured outcomes (e.g. BMI) that are likely associated with chronic rather than acute exposures. (Note that this evaluative criterion is not applicable to studies focused on exposure only, such as those examining temporal or spatial relationships within or across populations.)

3.2.2 Exposure variability and misclassification

For many short-lived chemicals, there can be large intra-individual temporal variability; attempting to find associations between one measure of such a chemical with disease is not supportable. Differences in biomonitored levels of short-lived chemicals due to changes in an individual's diet, health, product use, activity and/or location are expected (Pleil and Sobus, 2013). As noted by Meeker et al. (2013): "Characterizing temporal variability in exposure metrics, especially for biomarkers of nonpersistent compounds…, is a critical step in designing and interpreting an epidemiology study related to the potential for exposure measurement error."

Many published studies of short-lived chemicals seeking to estimate chronic or average exposure are subject to error because they rely on one measure of exposure using a one-time sample of

urine or blood (Goodman et al., 2014; LaKind et al., 2012b, 2014; Preau et al., 2010; Wielgomas, 2013). The ability to estimate exposure can be improved by taking multiple samples from the same individual at different times to average temporal variations in the biomarker levels (NRC, 2006). The reliability is typically measured by calculating the intra-class correlation coefficient (ICC). The ICC can be estimated by measuring the chemical in repeated samples collected over several hours, days or weeks and calculating the between-person variance divided by the total variance. ICCs range from 0 to 1; an ICC value equal to or approaching 1 suggests good reliability in estimating longer-term exposure for the population from a single sample. Symanski et al. (1996) used mixed-effects modeling to account for non-stationary behavior in occupational exposures, and found that estimates of variance components (used to compute ICC) may be substantially biased if systematic changes in exposure are not properly modeled. The following question still must be raised: if an ICC is developed from taking repeated samples over weeks or even months, will the value be relevant to exposures over years, which is the timeframe for development of many chronic diseases of interest? The research on this subject for many of the short-lived chemicals of interest is currently undeveloped.

Another problem with using a single measure of a short-lived chemical is error that may result in exposure misclassification. Exposure misclassification occurs when the assigned exposures do not correctly reflect the actual exposure levels or categories. It has been shown that exposure misclassification is difficult to predict in terms of both direction and magnitude (Cantor et al., 1992; Copeland et al., 1977; Dosemeci et al., 1990; Sorahan and Gilthorpe, 1994; Wacholder et al., 1995). The effect of exposure error and exposure misclassification on the dose-response relationship is problematic (Rhomberg et al., 2011). Exposure misclassification can occur from many sources of measurement error, including timing of sample collection relative to when a critical exposure occurs. For example, many volatile organic compounds have half-lives on the order of minutes; exposures may occur daily but for short time intervals. Thus, the concentration of the biomarker of exposure is highly dependent on when the sample is collected relative to when the exposure occurred and may not properly reflect the longer-term level in the body.

Use of multiple samples or prolonged (e.g., 24-hour) sample collection may help decrease error by diminishing the effects of temporal variation, study sub-population characteristics, and sample-related issues (Scher et al., 2007). If error cannot be avoided (e.g., if all available samples were obtained post-fast), it is important to assess accuracy of exposure characterization by calculating sensitivities and specificities (Jurek et al., 2006). Sensitivity is the probability of correctly classifying an individual as having high level of exposure, if that person truly belongs in the high exposure category. Specificity is the probability of correctly assigning low exposure to a participant who truly has a low level of exposure. Estimates of sensitivity and specificity may be calculated for a single urine sample, using multiple samples per subject as gold standard, since the true sensitivity and specificity for many measures is unknown. This can be achieved by randomly selecting a single sample from among each individual's repeated samples collected over the study (as demonstrated for phthalates in Adibi et al., 2008).

Short-lived chemical example: In a recent systematic review of the epidemiology literature on phthalates and associations with obesity, diabetes, and cardiovascular disease, Goodman et al. (2014) found that of 26 available studies, all but three relied on a single measure of phthalates. Similarly, in a systematic review of BPA and obesity, diabetes, and cardiovascular disease,

LaKind et al. (2014) found that of 45 available studies, all but four relied on a single measure of BPA. Yet the intra-individual variability for BPA is large (with ICCs ranging from 0.10 to 0.35 (Lassen et al., 2013; Teitelbaum et al., 2008), and multiple measures of exposure are needed to describe a person's long-term exposure. The ICCs for phthalates have been reported to be higher than for BPA (e.g., ICC values range from 0.18 to 0.61 for mono-ethyl phthalate, from 0.21 to 0.51 for mono-isobutyl phthalate, and from 0.08 to 0.27 for mono-(2-ethylhexyl) phthalate [reviewed in Goodman et al., 2014], but intra-person variability is still large. Recently, Attfield et al. (2014), in a study of variability of urinary pesticide measures in children, observed that a study with only a small number of samples from each study participant "…may lead to a high probability of exposure misclassification by incorrect quantile assignment and offer little assurance for correctly classifying the exposure into a specific category."

Study evaluation (Table 1): The above considerations permit dividing the available body of literature into the following tiers (Table 1). Tier 1 includes studies in which exposure assessment is based on sufficient number of samples per individual to estimate exposure over the appropriate duration, or through the use of adequate long-term sampling (e.g., multiple 24-hour urine collections). To be included in Tier 1, studies should assess error by calculating measures of accuracy (e.g., sensitivity and specificity) and reliability (e.g., ICC). It is possible that for some chemicals, one sample may be sufficient to fully characterize exposure. If this is the case, a Tier 1 study needs to provide evidence that errors of a single measurement can be considered sufficiently small. We realize this is not always feasible but there are circumstances where researcher will find it necessary to perform a validation study (Teeguarden et al. 2011). Tier 2 includes studies that use more than one sample, but provide no rationale for their choice of the number of measurements, and do not include an explicit evaluation of error. Tier 3 is reserved for studies in which exposure assessment is based on a single sample without considering error.

3.3 General Epidemiological Study Design Considerations

In this section, we discuss aspects of study design that are not necessarily specific to short-lived chemicals but are important in any assessment of overall study quality. Some of these issues are more applicable to those studies examining associations between exposure and health outcome while others may be applied to studies focused on exposure only.

3.3.1 Research Rationale

This section applies to hypothesis-testing studies examining associations between biomonitoring data and health outcome data. A well-formulated hypothesis arising from a clinical observation or from a basic science experiment is the cornerstone of any epidemiological inquiry regardless of the specific research field (Boet et al., 2012; Fisher and Wood, 2007; Moher and Tricco, 2008). Current recommendations in a variety of disciplines emphasize the importance of posing a research question that is structured to convey information about the population of interest, exposure (or corresponding marker) under investigation, and the outcome of concern (Sampson et al., 2009; Walker et al., 2012).

Biomonitoring studies – and in particular those involving short-lived chemicals where one sample can provide data on a multitude of chemicals - often generate data that contain multiple variables with an opportunity for multiple simultaneous hypothesis testing. This feature of biomonitoring studies can be viewed as a strength as in situations when significant associations

are observed for several related outcomes (Lord et al., 2004); e.g., if a hypothesized obesogen exerts similar effects on body mass index, waist circumference or percent body fat. On the other hand, the ability to assess multiple exposure-outcome associations complicates the interpretation of findings, particularly when dealing with previously collected data (Clarke et al., 2003; Lee and Huang, 2005; Marco and Larkin, 2000). Among studies that use previously collected data, it is important to distinguish those that were guided by an *a priori* formulated hypothesis from those that were conducted without a strong biological rationale, although the latter category has been proven helpful in formulating new hypotheses (Liekens et al., 2011; Oquendo et al., 2012). A study with a well-formulated hypothesis indicates that the study builds on previous knowledge, which is an important consideration for a WOE assessment. Studies specifically designed to add to the existing knowledge base can be more readily incorporated into WOE.

Study evaluation (Table 1): Studies evaluating an *a priori* formulated hypothesis with a biomonitoring strategy specifically designed to address this hypothesis should be considered the highest quality (Tier 1). Tier 2 studies would be those using existing samples or data to evaluate an *a priori* formulated hypothesis, where the biomonitoring strategy was not specifically designed for this purpose. In Tier 3 studies, the research relies on existing samples or data without a pre-specified hypothesis or involves multiple simultaneous hypothesis testing. We recognize that at present, the research rationale for most biomonitoring studies involving short-lived chemicals will be described as Tier 3 studies.

3.3.2 Study Participants

Evaluative schemes for participant selection apply to studies of both persistent and short-lived chemicals. The goal of participant selection in epidemiological research is to build a "bridge" between information that is obtainable from the sample and information sought about the target population (Kalsbeek and Heiss, 2000). The actual process of selecting an unbiased population sample is an ongoing challenge in case-control, longitudinal (cohort) and cross-sectional studies (Vandenbroucke et al., 2007).

The issue of participant selection is not unique to epidemiological research of short-lived chemicals. Yet biomonitoring studies may not pay sufficient attention to this problem. Previous reviews of biomonitoring studies presented evidence that selection bias may represent an important threat to internal validity (Bull et al., 2006; Faust et al., 2004). The same concerns are also applicable to biomonitoring studies of short-lived chemicals such as phthalates (Durmaz et al., 2010; Wang et al., 2013; Wirth et al., 2008).

Study evaluation (Table 1): Tier 1 studies include an unbiased selection and/or follow up protocol with a high (e.g., over 80%) response rate in cross-sectional or case-control studies, or low (e.g., less than 20%) loss to follow up in cohort studies. Tier 2 studies have an unbiased selection/follow up protocol and a low (e.g., 50%-80%) response rate in cross-sectional or case-control studies, or high (e.g., 20%-50%) loss to follow up in cohort studies. Tier 3 studies are those that include less than 50% of eligible participants, or fail to report methods of sample selection and/or rates of non-response or loss to follow up. A study that does not report this information should be assumed to be a Tier 3 study.

It is important to keep in mind that a low response rate or a high frequency of loss to follow-up should not be equated with selection bias. Selection bias occurs when the proportions of persons included in the final dataset (a.k.a. selection probabilities) differ by both exposure and outcome (e.g., among exposed cases, non-exposed cases, exposed non-cases and non-exposed non-cases.) Although the actual selection probabilities are usually unknown, one can expect that in a study that is missing only 10% of otherwise eligible participants, the magnitude of possible bias is much lower than the corresponding magnitude in a study that is missing 50% or more of its subjects.

3.3.3 Data Analysis

Essential aspects of data analysis in epidemiologic research have been reviewed elsewhere and are not specific to chemicals with short physiologic half lives. However, for completeness of the proposed tiered evaluative system, these considerations are described here in brief. The overall analytic strategy in observational research depends on the main goal of the study. Generally, statistical models fall into two categories – predictive and explanatory (Shmueli, 2010). For predictive analysis, selection of variables into the model is data-driven and may differ from dataset to dataset. The goal of this approach is to maximize the model fit and a decision on whether to retain a particular covariate of interest is based on statistical tests and goodness-of-fit without a specified exposure of interest (Bellazzi and Zupan, 2008). In an explanatory (hypothesis testing) analysis, this approach may be inappropriate because it may wrongly eliminate potentially important variables when the relationship between an outcome and a risk factor is confounded or may incorrectly retain variables that do not act as confounders (Kleinbaum and Klein 2002).

More importantly, for an explanatory model, which is focused on a pre-defined exposure-outcome association, inclusion and exclusion of control variables (confounders, mediators or effect modifiers) should be driven, at least in part, by *a priori* reasoning (Concato et al., 1993; Hernan et al., 2002; Beran and Violato, 2010).

It is important to keep in mind that the results of observational studies are inevitably subject to uncertainty. This uncertainty may be attributable to various sources of unaccounted bias and to various data handling decisions and assumptions. The magnitude of uncertainty can be formally assessed through quantitative sensitivity analyses. The methods of addressing residual bias through sensitivity analyses are now well developed both in terms of basic theory (Greenland, 1996) and with respect to practical applications (Goodman et al., 2007; Lash and Fink, 2003; Maldonado et al., 2003). With respect to sensitivity analyses of alternative decisions and assumptions, much can be learned from previous experience in economics, exposure assessment and quantitative risk analysis (Koornneef et al., 2010; Leamer, 1985; Spiegelman, 2010).

Study evaluation (Table 1): Tier 1 studies include those that clearly distinguish between causal and predictive models and demonstrate adequate consideration of extraneous factors with assessment of effect modification and adjustment for confounders. To qualify for Tier 1, a study should also perform formal sensitivity analyses. When consideration of extraneous factors is considered adequate and the model selection is appropriate, a study may still be considered incomplete without a sensitivity analysis. Those studies are placed in Tier 2. Tier 3 studies are

those that did not adequately control for extraneous factors due to inappropriate methods of covariate selection, failure to consider important confounders, or inability to take into account effect modification.

The term "extraneous factors" describes participant characteristics other than exposure and outcome of interest that need to be taken into consideration in the design or the analysis phase of the study because they may act as cofounders or effect modifiers or both (Kleinbaum et al. 2007).

3.3.4 Reporting of Results

We consider three aspects of reporting: transparency, multiple testing and reporting bias.

Reporting transparency: As noted in the STROBE statement, reporting of results should "ensure a clear presentation of what was planned, done, and found in an observational study" (Vandenbroucke et al., 2007). While these considerations are applicable to all studies, there are aspects of study reporting that are of particular relevance to biomonitoring research of short-lived chemicals.

Biological sample analyses are increasingly optimized for rapid analysis of multiple analytes in a single run. These developments in technology increase the importance of complete reporting of the data including a full list of exposure (and if applicable, outcome) biomarkers, as well as presentation of summary statistics, such as measures of central tendency and dispersion. Other critical information elements should include a description of patterns and handling of missing data and measures below LOD, all of which may influence interpretation of study results (Albert et al., 2010; Barnes et al., 2008; LaKind et al., 2012b). In addition, information should be provided on any power calculations used in determining the number of study participants and on the exposure gradient, which impacts the ability to identify significant associations. Although some of this information may not be included in the article due to space constraints, it can be incorporated in supplementary materials or made available upon request.

Considerations for multiple testing: The main concern with multiple hypothesis testing is increased likelihood of false positive (FP) results (Boffetta et al., 2008; Ioannidis, 2014; Jager and Leek, 2014; Rothman, 1990; Sabatti, 2007). Others have argued that a problem of FP results is no more important than the corresponding problem of false-negatives (FN) (Blair et al., 2009). A decision of what type of error (FP or FN) presents a greater concern is chemical- and outcome-specific, and should be made on a case-by-case basis. Recent advances in genetic and molecular epidemiology led to the development of novel approaches towards reducing the probability of FP (PFP) without increasing the risk of FN results (Datta and Datta, 2005; Wacholder et al., 2004). Even more recently, these approaches were further extended to allow calculating the FP:FN ratio (Ioannidis et al., 2011).

Reporting bias: When evaluating a body of research for a meta-analysis or WOE assessment, one must consider two specific sources of bias that may influence both analysis and synthesis of the available data: publication and outcome reporting bias. Publication bias is defined as the "tendency on the parts of investigators or editors to fail to publish study results on the basis of the direction or strength of the study findings" (Dickersin and Min, 1993). A closely related

concept is selective within-study reporting (a.k.a. outcome reporting bias), which is defined as "selection on the basis of the results of a subset of the original variables recorded for inclusion in a publication" (Dwan et al., 2008).

Publication bias is not specific to research involving short-lived chemicals. Outcome reporting bias, however, is potentially more problematic in studies of short-lived chemicals for reasons listed above. Specifically, better accessibility of sophisticated analytical platforms allows more analytes to be measured in a larger number of samples.

Study evaluation: A Tier 1 study clearly states its aims and allows the reader to evaluate the number of tested hypotheses (not just the number of hypotheses for which a result is given). If multiple simultaneous hypothesis testing is involved, its impact is assessed, preferably by estimating PFP or FP:FN ratio. There is no evidence of outcome reporting bias, and conclusions do not reach beyond the observed results. In a Tier 2 study, the conclusions appear warranted, but the number of tested hypotheses is unclear (either not explicitly stated or difficult to discern) and/or there is no consideration of multiple testing. Studies that selectively report data summaries and lack transparency in terms of methods or selection of presented results are included in Tier 3.

4. DISCUSSION/CONCLUSIONS

The need for a systematic approach to evaluating the quality of environmental epidemiology studies is clear. Two earlier efforts to develop evaluative schemes focused on epidemiology research on environmental chemical exposures and neurodevelopment (Amler et al., 2006; Youngstrom et al., 2011). Many of the concepts put forth in these proposed schemes are valuable to any evaluation of study quality and communicating study results when considering biomonitoring of chemicals with short physiologic half lives. For example, fundamental best practices/criteria proposed by Amler et al. (2006) include: a well-defined, biologically plausible hypothesis; the use of a prospective, longitudinal cohort design; consistency of research design protocols across studies; forthright, disciplined, and intellectually honest treatment of the extent to which results of any study are conclusive and generalizable; confinement of reporting to the actual research questions, how they were tested, and what the study found; recognition by investigators of their ethical duty to report negative as well as positive findings, and the importance of neither minimizing nor exaggerating these findings.

Chemicals with short physiologic half-lives present several important challenges, including their presence in analytical laboratories and sampling equipment, difficulty in establishing temporal order in cross-sectional studies, short- and long-term variability in exposures and biomarker concentrations, and a paucity of information on the number of measurements is required for accurate exposure classification. The BEES-C instrument is designed to evaluate these issues within a study or proposal.

We recognize that the development of an evaluative tool such as BEES-C is neither simple nor non-controversial, and we further expect that this will be an iterative process, similar to the data quality scheme that has been part of CONSORT and other existing methods or evaluating quality of clinical data. We also note that this type of evaluative scheme is not useful for exploratory

research; rather, the focus here is on designing and identifying those studies that have the greatest utility for furthering our understanding of associations between exposure to chemicals with short half lives and adverse health outcomes. We hope and anticipate that the instrument developed from this workshop will initiate further discussion/debate on this topic.

Acknowledgments: The views expressed in this publication were developed at a Workshop held in Baltimore Maryland in April, 2013. The Steering Committee included: Elaine Cohen Hubal, Ph.D., National Center for Computational Toxicology, U.S. EPA, Judy S. LaKind, Ph.D., LaKind Associates LLC, University of Maryland School of Medicine and Pennsylvania State University College of Medicine, Enrique F. Schisterman, Ph.D., Division of Epidemiology Statistics and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development National Institutes of Health, and Justin Teeguarden, PhD, DABT, Pacific Northwest National Laboratory. We thank three anonymous reviewers from the U.S. EPA and Health Canada for their thoughtful comments.

Competing Interests Declaration: The Workshop was sponsored by Polycarbonate/BPA Global Group of the American Chemistry Council (ACC). ACC was not involved in the design, management, or development of the Workshop or in the preparation or approval of the manuscript. Workshop participants or their affiliated organizations received an honorarium (except JSL, ES, GS, JS, JT, Y-MT, RT-V, TA) and travel support (except TA, Y-MT, DB, ES). JSL received support for Workshop development and facilitation; JSL consults to governmental and private sectors. MG regularly serves as a consultant for the government and for the private sector. No other competing interests are declared.

Disclaimer: The views expressed here are those of the authors and do not necessarily represent the views of the ACC, the US Environmental Protection Agency, Health Canada or the National Institute of Child Health and Human Development. The United States Environmental Protection Agency through its Office of Research and Development collaborated in the research described here. It has been subjected to Agency review and approved for publication.

5. REFERENCES

Adibi JJ, Whyatt RM, Williams PL, Calafat AM, Camann D, Herrick R, et al. 2008. Characterization of phthalate exposure among pregnant women assessed by repeat air and urine samples. Environ Health Perspect 116:467–473.

Albert PS, Harel O, Perkins N, Browne R. 2010. Use of multiple assays subject to detection limits with regression modeling in assessing the relationship between exposure and outcome. Epidemiology 21 Suppl 4:S35–43.

Amler RW, Barone S, Jr., Belger A, Berlin CM, Jr., Cox C, Frank H, et al. 2006. Hershey Medical Center Technical Workshop Report: Optimizing the design and interpretation of epidemiologic studies for assessing neurodevelopmental effects from in utero chemical exposure. Neurotoxicology 27:861–874.

Ankley GT, Bennett RS, Erickson RJ, Hoff DJ, Hornung MW, Johnson RD, et al. 2010. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environ Toxicol Chem 29:730–41.

Attfield KR, Hughes MD, Spengler JD, Lu C. 2014. Within- and between-child variation in repeated urinary pesticide metabolite measurements over a 1-year period. Environ Health Perspect 122:201–206.

Barnes SA, Mallinckrodt CH, Lindborg SR, Carter MK. 2008. The impact of missing data and how it is handled on the rate of false-positive results in drug development. Pharm Stat 7:215–225.

Barr DB, Barr JR, Driskell WJ, Hill RH, Jr., Ashley DL, Needham LL, et al. 1999. Strategies for biological monitoring of exposure for contemporary-use pesticides. Toxicol Ind Health 15:168–179.

Barr DB, Landsittel D, Nishioka M, Thomas K, Curwin B, Raymer J, et al. 2006. A survey of laboratory and statistical issues related to farmworker exposure studies. Environ Health Perspect 114:961–968.

Barr DB, Needham LL. 2002. Analytical methods for biological monitoring of exposure to pesticides: a review. J Chromatogr B Analyt Technol Biomed Life Sci 778:5–29.

Barr DB, Olsson AO, Wong LY, Udunka S, Baker SE, Whitehead RD, et al. 2010. Urinary concentrations of metabolites of pyrethroid insecticides in the general U.S. population: National Health and Nutrition Examination Survey 1999-2002. Environ Health Perspect 118:742–748.

Barr DB, Wang RY, Needham LL. 2005a. Biologic monitoring of exposure to environmental chemicals throughout the life stages: requirements and issues for consideration for the National Children's Study. Environ Health Perspect 113:1083–1091.

Barr DB, Wilder LC, Caudill SP, Gonzalez AJ, Needham LL, Pirkle JL. 2005b. Urinary creatinine concentrations in the U.S. population: implications for urinary biologic monitoring measurements. Environ Health Perspect 113:192–200.

Bellazzi R, Zupan B. 2008. Predictive data mining in clinical medicine: Current issues and guidelines. Int J Med Inform 77:81–97.

Beran TN, Violato C. 2010. Structural equation modeling in medical research: A primer. BMC Res Notes 3:267.

Blair A, Saracci R, Vineis P, Cocco P, Forastiere F, Grandjean P, et al. 2009. Epidemiology, public health, and the rhetoric of false positives. Environ Health Perspect 117:1809–1813.

Boet S, Sharma S, Goldman J, Reeves S. 2012. Review article: Medical education research: An overview of methods. Can J Anaesth 59:159–170.

Boffetta P, McLaughlin JK, La Vecchia C, Tarone RE, Lipworth L, Blot WJ. 2008. False-positive results in cancer epidemiology: A plea for epistemological modesty. J Natl Cancer Inst 100:988–995.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al.; STARD Group. 2004. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. Fam Pract 21:4–10.

Bouchard M, Krishnan K, Viau C. 1998. Kinetics of tissue distribution and elimination of pyrene and 1-hydroxypyrene following intravenous administration of [14C]pyrene in rats. Toxicol Sci 46:11–20.

Bouchard M, Viau C. 1997. Urinary excretion of benzo[a]pyrene metabolites following intravenous, oral, and cutaneous benzo[a]pyrene administration. Can J Physiol Pharmacol 75:185–192.

Bull S, Fletcher K, Boobis AR, Battershill JM. 2006. Evidence for genotoxicity of pesticides in pesticide applicators: A review. Mutagenesis 21:93–103.

Calafat AM, Needham LL. 2008. Factors affecting the evaluation of biomonitoring data for human exposure assessment. Int J Androl 31:139–143.

Calafat AM, Needham LL. 2009. What additional factors beyond state-of-the-art analytical methods are needed for optimal generation and interpretation of biomonitoring data? Environ Health Perspect 117:1481–1485.

Calafat AM, Koch HM, Swan SH, Hauser R, Goldman LR, Lanphear BP, et al. 2013. Misuse of blood serum to assess exposure to bisphenol A and phthalates. Breast Cancer Res 15:403.

Cantor KP, Blair A, Everett G, Gibson R, Burmeister LF, Brown LM, et al. 1992. Pesticides and other agricultural risk factors for non-Hodgkin's lymphoma among men in Iowa and Minnesota. Cancer Res 52:2447–2455.

Caudill SP. 2010. Characterizing populations of individuals using pooled samples. J Expo Sci Environ Epidemiol 20:29–37.

Clarke P, Sproston K, Thomas R. 2003. An investigation into expectation-led interviewer effects in health surveys. Social Sci Med 56:2221–2228.

Concato J, Feinstein AR, Holford TR. 1993. The risk of determining risk with multivariable models. Ann Intern Med 118:201–210.

Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. 1977. Bias due to misclassification in the estimation of relative risk. Am J Epidemiol 105:488–495.

Datta S, Datta S. 2005. Empirical Bayes screening of many p-values with applications to microarray studies. Bioinformatics 21:1987–1994.

Dickersin K, Min YI. 1993. Publication bias: The problem that won't go away. Ann N Y Acad Sci 703:135–146; discussion 146–138.

Dosemeci M, Wacholder S, Lubin JH. 1990. Does nondifferential misclassification of exposure always bias a true effect toward the null value? Am J Epidemiol 132:746–748.

Durmaz E, Ozmert EN, Erkekoglu P, Giray B, Derman O, Hincal F, et al. 2010. Plasma phthalate levels in pubertal gynecomastia. Pediatrics 125:e122–9.

Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, et al. 2008. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS One 3:e3081.

EPA (US Environmental Protection Agency). 2013. America’s Children and the Environment. Third Edition. Available: http://www.epa.gov/ace/ Accessed November 25, 2013.

Faust F, Kassie F, Knasmuller S, Boedecker RH, Mann M, Mersch-Sundermann V. 2004. The use of the alkaline comet assay with lymphocytes in human biomonitoring studies. Mutat Res 566:209–229.

Fisher CG, Wood KB. 2007. Introduction to and techniques of evidence-based medicine. Spine (Phila Pa 1976) 32:S66–72.

Gallo V, Egger M, McCormack V, Farmer PB, Ioannidis JPA, Kirsch-Volders M, Matullo G, Phillips DH, Schoket B, Stromberg U, Vermeulen R. 2011. STrengthening the Reporting of OBservational studies in Epidemiology Molecular Epidemiology STROBE-ME: an extension of the STROBE statement. J Clin Epidemiol 64:1350–1363.

Goodman M, Barraj LM, Mink PJ, Britton NL, Yager JW, Flanders WD, et al. 2007. Estimating uncertainty in observational studies of associations between continuous variables: Example of methylmercury and neuropsychological testing in children. Epidemiol Perspect Innov 4:9.

Goodman M, LaKind JS, Mattison DR. 2014. Do phthalates act as obesogens in humans? A systematic review of the epidemiology literature. Crit Rev Toxicol In press.

Gordis L. 2008. Epidemiology. Philadelphia, PA: Saunders Elsevier.

Greenland S. 1996. Basic methods for sensitivity analysis of biases. Int J Epidemiol 25:1107–1116.

Harthé C, Rinaldi S, Achaintre D, de Ravel MR, Mappus E, Pugeat M, Déchaud H. 2012. Bisphenol A-glucuronide measurement in urine samples. Talanta 100:410-413.

Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. 2002. Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. Am J Epidemiol 155:176–184.

Hertz-Picciotto I. Environmental Epidemiology. 1998. In: Modern Epidemiology Rothman KJ, Greenland S. (Eds.) Lippincott Williams and Wilkins.

Hytten F. 1985. Blood volume changes in normal pregnancy. Clin Haematol 14:601–12.

Ioannidis JP. 2014. Discussion: Why "an estimate of the science-wise false discovery rate and application to the top medical literature" is false. Biostatistics 15:28–36; discussion 39–45.

Ioannidis JP, Tarone R, McLaughlin JK. 2011. The false-positive to false-negative ratio in epidemiologic studies. Epidemiology 22:450–456.

Jager LR, Leek JT. 2014. An estimate of the science-wise false discovery rate and application to the top medical literature. Biostatistics 15:1–12.

Jurek AM, Maldonado G, Greenland S, Church TR. 2006. Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. Eur J Epidemiol 21:871–876.

Kalsbeek W, Heiss G. 2000. Building bridges between populations and samples in epidemiological studies. Annu Rev Public Health 21:147–169.

Kissel JC, Curl CL, Kedan G, Lu C, Griffith W, Barr DB, et al. 2005. Comparison of organophosphorus pesticide metabolite levels in single and multiple daily urine samples collected from preschool children in Washington State. J Expo Anal Environ Epidemiol 15:164–171.

Kleinbaum DG. Klein M. 2002. Logistic Regression: A Self-Learning Text. Springer-Verlag New York, NY.

Kleinbaum DG, Sullivan KM, Barker ND. 2007. A Pocket Guide to Epidemiology. Springer Science + Business Media:New York pp. 228–229.

Koornneef J, Spruijt M, Molag M, Ramirez A, Turkenburg W, Faaij A. 2010. Quantitative risk assessment of co2 transport by pipelines--a review of uncertainties and their impacts. J Hazard Mater 177:12–27.

Kuo CC, Moon K, Thayer KA, Navas-Acien A. 2013. Environmental chemicals and type 2 diabetes: an updated systematic review of the epidemiologic evidence. Curr Diab Rep 13:831–49.

LaKind JS, Naiman DQ. 2008. Bisphenol A (BPA) daily intakes in the United States: Estimates from the 2003-2004 NHANES urinary BPA data. J Exp Sci Environ Epidemiol 18:608–615.

LaKind JS, Naiman DQ. 2011. Daily intake of bisphenol A (BPA) and potential sources of exposure – 2005-2006 NHANES. J Exp Sci Environ Epidemiol 21:272–279.

LaKind JS, Levesque J, Dumas P, Bryan S, Clarke J, Naiman DQ. 2012a. Comparing United States and Canadian population exposures from national biomonitoring surveys: Bisphenol A intake as a case study. J Exp Sci Environ Epidemiol 22:219–226.

LaKind JS, Goodman M, Naiman DQ. 2012b. Use of NHANES data to link chemical exposures to chronic diseases: a cautionary tale. PLoS ONE 7(12):e51086. doi:10.1371/journal.pone.0051086

LaKind JS, Goodman M, Mattison DR. 2014. Bisphenol A and indicators of obesity, glucose metabolism/type 2 diabetes and cardiovascular disease: A systematic review of epidemiologic research. Crit Rev Toxicol In press.

Lanphear BP, Dietrich K, Auinger P, Cox C. 2000. Cognitive deficits associated with blood lead concentrations <10 microg/dL in US children and adolescents. Public Health Rep 115:521–529.

Lash TL, Fink AK. 2003. Semi-automated sensitivity analysis to assess systematic errors in observational data. Epidemiology 14:451–458.

Lassen TH, Frederiksen H, Jensen TK, Petersen JH, Main KM, Skakkebæk NE, et al. 2013. Temporal variability in urinary excretion of bisphenol A and seven other phenols in spot, morning, and 24-h urine samples. Env Res 126:164–70.

Lee WC, Huang HY. 2005. Data-dredging gene-dose analyses in association studies: Biases and their corrections. Cancer Epidemiol Biomarkers Prev 14:3004–3006.

Leamer EE. 1985. Sensitivity analyses would help. Am Econ Rev 75:308-313.

Leng G, Kuhn KH, Idel H. 1997. Biological monitoring of pyrethroids in blood and pyrethroid metabolites in urine: applications and limitations. Sci Total Environ 199:173-181.

Liekens AM, De Knijf J, Daelemans W, Goethals B, De Rijk P, Del-Favero J. 2011. Biograph: Unsupervised biomedical knowledge discovery via automated hypothesis generation. Genome Biol 12:R57.

Little J, Higgins JP, Ioannidis JP, Moher D, Gagnon F, von Elm E, et al. 2009. Strengthening the reporting of genetic association studies (STREGA): An extension of the strengthening the reporting of observational studies in epidemiology (STROBE) statement. J Clin Epidemiol 62:597–608 e594.

Lorber M, Koch HM, Angerer J. 2011. A critical evaluation of the creatinine correction approach: can it underestimate intakes of phthalates? A case study with di-2-ethylhexyl phthalate. J Expo Sci Environ Epidemiol 21:576–586.

Lord SJ, Gebski VJ, Keech AC. 2004. Multiple analyses in clinical trials: Sound science or data dredging? Med J Aust 181:452–454.

Maldonado G, Delzell E, Tyl RW, Sever LE. 2003. Occupational exposure to glycol ethers and human congenital malformations. Int Arch Occup Environ Health 76:405–423.

Marco CA, Larkin GL. 2000. Research ethics: Ethical issues of data reporting and the quest for authenticity. Acad Emerg Med 7:691–694.

Markham DA, Waechter JM Jr, Wimber M, Rao N, Connolly P, Chuang JC, et al. 2010. Development of a method for the determination of bisphenol A at trace concentrations in human blood and urine and elucidation of factors influencing method accuracy and sensitivity. J Anal Toxicol 34:293–303.

Meeker JD, Barr DB, Ryan L, Herrick RF, Bennett DH, Bravo R, et al. 2005. Temporal variability of urinary levels of nonpersistent insecticides in adult men. J Expo Anal Environ Epidemiol 15:271–281.

Meeker JD, Cantonwine DE, Rivera-González LO, Ferguson KK, Mukherjee B, Calafat AM, et al. 2013. Distribution, variability, and predictors of urinary concentrations of phenols and parabens among pregnant women in Puerto Rico. Environ Sci Technol 47:3439–3447.

Moher D, Schulz KF, Altman DG. 2001. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 357:1191–1194.

Moher D, Tricco AC. 2008. Issues related to the conduct of systematic reviews: A focus on the nutrition field. Am J Clin Nutr 88:1191–1199.

National Research Council (NRC). 2006. Human Biomonitoring for Environmental Chemicals. Washington, DC: The National Academies Press.

National Toxicology Program (NTP). 2013. Draft OHAT Approach For Systematic Review and Evidence Integration For Literature-Based Health Assessments – February. Division of the National Toxicology Program, National Institute of Environmental Health Sciences, National Institutes of Health. Available: http://ntp.niehs.nih.gov/?objectid=960B6F03-A712-90CB-8856221E90EDA46E [accessed 25 October 2013]

Needham LL, Calafat AM, Barr DB. 2007. Uses and issues of biomonitoring. Int J Hyg Environ Health 210:229–238.

Oquendo MA, Baca-Garcia E, Artes-Rodriguez A, Perez-Cruz F, Galfalvy HC, Blasco-Fontecilla H, et al. 2012. Machine learning and data mining: Strategies for hypothesis generation. Mol Psychiatry 17:956–959.

Owens DK, Lohr KN, Atkins D, Treadwell JR, Reston JT, Bass EB, et al. 2010. AHRQ series paper 5: Grading the strength of a body of evidence when comparing medical interventions--Agency for Healthcare Research and Quality and the effective health-care program. J Clin Epidemiol 63:513–523.

Pearce N. 2012. Classification of epidemiological study designs. Int J Epidemiol 41:393–397.

Pleil JD, Sobus JR. 2013. Estimating lifetime risk from spot biomarker data and intraclass correlation coefficients (ICC). J Toxicol Environ Health, Part A 76:747–766.

Potischman N, Weed DL. 1999. Causal criteria in nutritional epidemiology. Am J Clin Nutr 69:1309S–1314S.

Preau JL Jr, Wong LY, Silva MJ, Needham LL, Calafat AM. 2010. Variability over 1 week in the urinary concentrations of metabolites of diethyl phthalate and di(2-ethylhexyl) phthalate among eight adults: an observational study. Environ Health Perspect 118:1748–1754.

Rappaport SM, Symanski E, Yager JW, Kupper LL. 1995. The relationship between environmental monitoring and biological markers in exposure assessment. Environ Health Perspect 103(Suppl 3):49–53.

Rhomberg LR, Chandalia JK, Long CM, Goodman JE. 2011. Measurement error in environmental epidemiology and the shape of exposure-response curves. Crit Rev Toxicol 41:651–671.

Rothman KJ. 1990. No adjustments are needed for multiple comparisons. Epidemiology 1:43–46.

Rothman KJ, Greenland S. 1998. Modern epidemiology. Philadelphia, PA:Lippincott Williams and Wilkins.

Rothman KJ, Greenland S. 2005. Causation and causal inference in epidemiology. Am J Public Health 95 Suppl 1:S144–150.

Sabatti C. 2007. Avoiding false discoveries in association studies. Meth Mol Biol 376:195–211.

Sampson M, McGowan J, Cogo E, Grimshaw J, Moher D, Lefebvre C. 2009. An evidence-based practice guideline for the peer review of electronic search strategies. J Clin Epidemiol 62:944–952.

Scher DP, Alexander BH, Adgate JL, Eberly LE, Mandel JS, Acquavella JF, et al. 2007. Agreement of pesticide biomarkers between morning void and 24-h urine samples from farmers and their children. J Expo Sci Environ Epidemiol 17:350–357.

Schisterman EF, Whitcomb BW, Buck Louis GM, Louis TA. 2005. Lipid Adjustment in the Analysis of Environmental Contaminants and Human Health Risks. Environ Health Perspect 113:853–857.

Shmueli G. 2010. To explain or to predict? Stat Sci 25:289–310.

Sobus JR, McClean MD, Herrick RF, Waidyanatha S, Nylander-French LA, Kupper LL, et al. 2009. Comparing urinary biomarkers of airborne and dermal exposure to polycyclic aromatic compounds in asphalt-exposed workers. Ann Occup Hyg 53:561–571.

Sorahan T, Gilthorpe MS. 1994. Non-differential misclassification of exposure always leads to an underestimate of risk: an incorrect conclusion. Occup Environ Med 51:839–840.

Spiegelman D. 2010. Approaches to uncertainty in exposure assessment in environmental epidemiology. Annu Rev Public Health 31:149–163.

Symanski E, Kupper LL, Kromhout H, Rappaport SM. 1996. An investigation of systematic changes in occupational exposure. 57:724–35.

Teeguarden JG, Calafat AM, Ye X, Doerge DR, Churchwell MI, Gunawan R, et al. 2011. Twenty-four hour human urine and serum profiles of bisphenol a during high-dietary exposure. Toxicol Sci 123:48–57.

Teitelbaum SL, Britton JA, Calafat AM, Ye X, Silva MJ, Reidy JA, et al. 2008. Temporal variability in urinary concentrations of phthalate metabolites, phytoestrogens and phenols among minority children in the United States. Environ Res 106:257–269.

Vandenbroucke JP, von Elm E, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ, et al. 2007. Strengthening the reporting of observational studies in epidemiology (strobe): Explanation and elaboration. Epidemiology 18:805–835.

Viau C, Bouchard M, Carrier G, Brunet R, Krishnan K. 1999. The toxicokinetics of pyrene and its metabolites in rats. Toxicol Lett 108:201–207.

Völkel W, Kiranoglu M, Fromme H. 2008. Determination of free and total bisphenol A in human urine to assess daily uptake as a basis for a valid risk assessment. Toxicol Lett. 179:155-162.

Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. 2004. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. J Natl Cancer Inst 96:434-442.

Wacholder S, Hartge P, Lubin JH, Dosemeci M. 1995. Non-differential misclassification and bias towards the null: a clarification. Occup Environ Med 52:557-558.

Walker DG, Wilson RF, Sharma R, Bridges J, Niessen L, Bass EB, et al. 2012. Best practices for conducting economic evaluations in health care: A systematic review of quality assessment tools. (AHRQ Methods for Effective Health Care). Rockville (MD):Agency for Healthcare Research and Quality.

Wang H, Zhou Y, Tang C, He Y, Wu J, Chen Y, Jiang Q. 2013. Urinary phthalate metabolites are associated with body mass index and waist circumference in Chinese school children. PLoS One 8:e56800.

Weed DL, Gorelic LS. 1996. The practice of causal inference in cancer epidemiology. Cancer Epidemiol Biomarkers Prev 5:303-311.

Weed DL. 1997. On the use of causal criteria. Int J Epidemiol 26:1137–1141.

Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. 2003. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 3:25.

WHO (World Health Organization). 2011. Toxicological and Health Aspects of Bisphenol A. Report of Joint FAO/WHO Expert Meeting. 2–5 November 2010 and Report of Stakeholder Meeting on Bisphenol A. Available: whqlibdoc.who.int/publications/2011/97892141564274_eng.pdf [accessed 25 November 2013]

Wielgomas B. 2013. Variability of urinary excretion of pyrethroid metabolites in seven persons over seven consecutive days-Implications for observational studies. Toxicol Lett 221:15–22.

Wirth JJ, Rossano MG, Potter R, Puscheck E, Daly DC, Paneth N, et al. 2008. A pilot study associating urinary concentrations of phthalate metabolites and semen quality. Syst Biol Reprod Med 54:143–154.

Withey JR, Law FC, Endrenyi L. 1991. Pharmacokinetics and bioavailability of pyrene in the rat. J Toxicol Environ Health 32:429–447.

Ye X, Kuklenyik Z, Needham LL, Calafat AM. 2005. Quantification of urinary conjugates of bisphenol A, 2,5-dichlorophenol, and 2-hydroxy-4-methoxybenzophenone in humans by online solid phase extraction-high performance liquid chromatography-tandem mass spectrometry. Anal Bioanal Chem 383:638-644.

Ye X, Tao LJ, Needham LL, Calafat AM. 2008. Automated on-line column-switching HPLC-MS/MS method for measuring environmental phenols and parabens in serum. Talanta 76:865–871.

Ye X, Zhou X, Hennings R, Kramer J, Calafat AM. 2013. Potential external contamination with bisphenol A and other ubiquitous organic environmental chemicals during biomonitoring analysis: An elusive laboratory challenge. Environ Health Perspect 121:283–286.

Youngstrom E, Kenworthy L, Lipkin PH, Goodman M, Squibb K, Mattison DR, et al. 2011. A proposal to facilitate weight-of-evidence assessments: Harmonization of Neurodevelopmental Environmental Epidemiology Studies (HONEES). Neurotoxicol Teratol 33:354–359.

Zartarian V, Bahadori T, McKone T. 2005. Adoption of an official ISEA glossary. J Exp Anal Environ Epidemiol 15:1–5.

Zelenka MP, Barr DB, Nicolich MJ, Lewis RJ, Bird MG, Letinski DJ, et al. 2011. A weight of evidence approach for selecting exposure biomarkers for biomonitoring. Biomarkers 16:65–73.

Zota AR, Calafat AM, Woodruff TJ. 2014. Temporal trends in phthalate exposures: findings from the National Health and Nutrition Examination Survey, 2001-2010. Environ Health Perspect 122:235–241.