Biomarkers, 2011; 16(1): 65–73

informa
healthcare

ORIGINAL ARTICLE

A weight of evidence approach for selecting exposure biomarkers for biomonitoring

Michael P. Zelenka[1]*, Dana Boyd Barr[2], Mark J. Nicolich[3], R. Jeffrey Lewis[1], Michael G. Bird[1], Daniel J. Letinski[1], Susan W. Metcalf[4], Raegan B. O'Lone[5]

[1]ExxonMobil Biomedical Sciences, Inc., Annandale, New Jersey, USA, [2]Rollins School of Public Health, Emory University, Atlanta, Georgia, USA, [3]COGIMET, Lambertville, New Jersey, USA, [4]Agency for Toxic Substances and Disease Registry, Atlanta, Georgia, USA, and [5]ILSI Health and Environmental Sciences Institute, Washington, D.C., USA

Abstract

Context: It is known that there are usually several biomarkers and/or medium combinations that can be applied to answer a specific exposure question. To help determine an appropriate combination for the specific question, we have developed a weight-of-evidence Framework that provides a relative appropriateness score for competing combinations.

Methods: The Framework is based on an expert assessor's evaluation of the relevance and suitability of the biomarker and medium for the question based on a set of criteria. We provide a computer based modeling tool to guide the researcher through the process.

Results: We present an example with six biomarkers of benzene exposure in one matrix; the six are either the most commonly used biomarkers and/or have recent widespread usage. The example clearly demonstrates the usefulness of the Framework for scoring the choices, as well as the transparency of the method that provides the basis for discussion.

Conclusions: The Framework provides for the first time a method to transparently document the rationale behind selecting, from among a set of alternatives, the most scientifically supportable exposure biomarker to address a specific biomonitoring question, thus providing a reproducible account of expert opinions on the suitability of a biomarker.

Keywords: *Biological monitoring; biomarker of exposure; weight of evidence calculator; benzene*

Abbreviations: ACGIH: American Conference of Governmental Industrial Hygienists , ATSDR: Agency for Toxic Substances and Disease Registry, DFG: Deutsche Forschungsgemeinshaft, DNA: deoxyribonucleic acid, ECETOC: European Centre for Ecotoxicology and Toxicology of Chemicals, ILSI-HESI: International Life Sciences Institute - Health and Environmental Sciences Institute, NRC: National Research Council, ppm: parts per million, s-PMA: s-phenylmercapturic acid, tt-MA: trans-trans muconic acid, TWA: time-weighted average, WHO: World Health Organization, WoEC: Weight of Evidence Calculator

Introduction

Biological monitoring, or biomonitoring, has developed rapidly over the past two decades. Biomonitoring typically consists of taking samples from a subject (usually exhaled breath, urine, blood, hair, or adipose tissue) to measure trace levels of chemicals or metabolites of chemicals in the sample. The National Academy of Sciences defines biomonitoring as, "one method for assessing human exposure to chemicals by measuring the chemicals or their metabolites in human tissues or specimens, such as blood and urine" (NRC 2006). The concentration of a specific chemical in a biological sample may reflect a person's exposure to the chemical. Biological markers, or biomarkers, can be an environmental substance, a metabolite of the substance, or a

Address for Correspondence: ExxonMobil Biomedical Sciences, Inc. 1545 US Highway 22 East, Room LA388, Annandale, NJ 08801-3059, Telephone: 908-730-1066, Fax: 908-730-1192, E-mail: michael.p.zelenka@exxonmobil.com

(Received 01 July 2010; revised 06 October 2010; accepted 16 October 2010)

ISSN 1354-750X print/ISSN 1366-5804 online © 2011 Informa UK, Ltd.
DOI: 10.3109/1354750X.2010.533283

product (e.g. adducts) of its reaction with a protein, nucleic acid, organelle, or a biological process that may be measured in human tissue or body fluids (WHO 2001, ECETOC 2005, Bird 2008). A valid biomarker of exposure is one that links the biomarker in the subject to a specific environmental exposure. Various endpoints can serve as biological markers in biological systems or samples. In some cases, these molecular or cellular effects (e.g. DNA or protein adducts, mutations, chromosomal aberrations, levels of thyroid-stimulating hormone) that can be measured in blood, body fluids, cells, and tissues may serve as biomarkers of exposure in both humans and animals. For the purposes of this paper, the discussion of biological markers will be limited to biomarkers of human exposure.

Biological monitoring affords the ability to assess human exposure to chemicals by all routes, including inhalation, ingestion, and dermal absorption. Selecting an appropriate biomarker for an exposure requires knowledge of the distribution, metabolism, and excretion of the substance sufficient for selection of the proper metabolite to be determined, biological medium (or biological matrix) to be sampled, and time constraint for obtaining a specimen. Ideally, a biomarker will be unique to the substance under investigation, but this is often not the case.

Analytical techniques have made detection of ever smaller quantities of a biomarker possible generating increased public awareness of the body burden of chemicals. This highlights the need for identification of valid biomarkers for specific purposes. It is well known that there are several biomarker(s) and/or medium combinations that can be used to assess a specific exposure. Other organizations and expert bodies have reviewed biomarker case studies and noted the need for a framework to optimize the selection from among potential biomarkers (ECETOC 2005, Albertini et al. 2006, NRC 2006, ILSI-HESI 2008). For example, when estimating benzene exposure, phenol has been the traditional biomarker that is valid for higher exposures. However, at lower benzene exposures, phenol becomes less valid as a biomarker because phenol exposure also occurs at lower levels through one's diet and it is not possible to distinguish ingested phenol from benzene metabolized phenol. This makes it unreliable as a biomarker when inhalation exposures are low (i.e. less than about 10 ppm) (ACGIH 2001). Therefore, benzene estimation at lower exposure levels should be based on other biomarkers which are more specific and sensitive such as benzene in blood, or urinary *S*-phenylmercapturic acid. It is therefore important to use the appropriate biomarker for a specified exposure scenario taking into account the estimated external exposure level, and other equally important variables such as the biological material to be analyzed, the time frame from exposure to measurement, ease of obtaining a sample, and other considerations.

A European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC) Task Force and a National Research Council committee developing guidance for the interpretation of biomonitoring data, expressed the need for a framework to help identify the particular biomarker that is best suited for use in a particular circumstance (ECETOC 2005, NRC 2006). In response to the stated need, we developed a weight-of-evidence based Framework for determining the appropriateness of a biomarker for specific questions of exposure. The development was sponsored by the International Life Sciences Institute Health and Environmental Science Institute's (ILSI-HESI) Integration of Biomonitoring Exposure Data into the Risk Assessment Process (Biomonitoring) Technical Committee. The mission of the technical committee is to delineate the appropriate scientific uses of biomonitoring tools and biomonitoring data needed to characterize exposure to chemicals, and to explore mechanisms for integrating biomonitoring and toxicology data into a robust risk assessment process (Doerrer 2007). The Framework is based on the interpretive criteria developed for the International Biomonitoring Workshop co-sponsored by ILSI-HESI in 2004 (Albertini et al. 2006, Bird 2008). The criteria cover a wide range of aspects of a biomarker's properties. For example, one of the criteria is, "Are the pharmacokinetics [of the biomarker] well understood?" Table 1 lists the criteria available to be evaluated for a biomarker of exposure. Each of these criteria can be evaluated based on their relationship to the biomarker being assessed. The criteria form a basis for the assessor to determine the most scientifically supportable biomarker for a given question of exposure. The criteria and candidate biomarker characteristics are rated on a numerical scale that is related to their importance and suitability to the specific question of exposure. The resulting numerical scores are summarized to provide a relative ranking of the suitability of each candidate biomarker to the exposure question. An individual criterion is described by broad categories based in part on an international biomonitoring workshop that recommended criteria for applying and interpreting biomonitoring information (Albertini et al. 2006). The inclusion and location of each criterion within the broad categories is presented in this form as a flexible framework.

The Framework serves three useful purposes:

1. It can be used to make a transparent, reproducible account of expert opinions on the suitability of a biomarker.
2. It provides a method of scoring the biomarkers based on the experts' recorded opinions. The opinions are catalogued through individual criterion items.

Table 1. Listing of interpretive criteria in WoEC.

Criteria Sections	Criteria
Exposure/Toxicology	Can the source(s) be determined for the biomarkers?
	Can the transport medium be determined for the biomarkers?
	Can the exposure route(s) be determined for the biomarkers?
	Do historical biomarker data exist?
	Are longitudinal data important?
	Does one need to consider the ease of acquiring a body sample?
	Are the pharmacokinetics well understood?
	Is the inter-individual variation well understood?
	Is knowing the half-life of the biomarker in the body important?
	Is it important to be able to distinguish a single exposure?
	Is it important to be able to distinguish between multiple exposures?
	Is it important that the concentration of the biomarker increase with exposure?
	Is it necessary to distinguish between sources?
	Is the cost per sample important?
	Is the cost of a study important?
Sampling	Is the biomarker specific to the exposure?
	Are pre-analytic and analytic contamination of concern?
	Is the biomarker stable in the sample matrix?
	Do you need to have multiple analytic lab ability to analyze the biomarker?
Analytic Methodology	Does an analytic method exist to measure the biomarker?
	Are standard materials available?
	Are internal standards available?
	Is a quantitative measure necessary?
	Is a confirmatory analysis necessary?
	Is the sensitivity (i.e. method limit of detection) important?
	Is the precision of the method important?
	Is it important to have certified (i.e. traceable) accuracy standards?
	Is the method rugged?
	Does the method have an appropriate quality assurance/quality control system?
	Is having an external QC system necessary?
	Does having laboratory certification affect the quality of the data?
	Is having inter-laboratory comparisons important?

3. It assists researchers in identifying key data gaps in their knowledge about specific biomarkers of exposure particularly as they relate to pre-defined applications of the biomarkers.

In summary, the Framework is an aid in the selection of biomarkers of exposure for a specific purpose, based on a set of detailed criteria that builds upon the ECETOC and NRC reports (ECETOC 2005, NRC 2006).

Methods

The Framework is based on developing a summary score for a biomarker to address a specific goal. The score incorporates the *relevance* of a criterion for meeting the goal and the *suitability* of the biomarker to address the criterion. The final score is a summary of the contributions of all the criteria included in an analysis. The method gives positive weight to a criterion that is important and a biomarker that achieves the criterion, it gives negative weight to an important criterion that the biomarker does not address, and gives neutral weight to an unimportant criterion no matter how the biomarker addresses the criterion. This method has been successfully applied in determining consumer preferences (Roy and Nicolich 1980), it is the basis for the Analytic Hierarchy Process (Bhushan and Rai 2004), and shares some characteristics with modern decision analysis (Clemen 1996).

The application of the Framework to a specific problem requires a series of steps that will be outlined in the following section. These steps have been incorporated into an Excel spreadsheet to guide the researcher and perform the necessary calculations. An example of the application of the Framework to a specific problem and the use of the spreadsheet are provided in the Results section.

A.Framework description

The evaluation process is comprised of six steps, all of which are important for the assessment and should

be completed. What follows is an outline of the six steps followed by a description of each of the six steps in which the term 'researcher' applies to the group or expert conducting the analysis. The six steps are:

1. specify the intended use of the biomarker,
2. rigorously define the exposure question against which the biomarkers will be evaluated,
3. develop the list of competing biomarkers and matrices,
4. rate each criterion as to its relevance in answering the exposure question,
5. rate the suitability of each biomarker for each criterion,
6. complete the required calculations and interpretation.

The first step is to specify the intended use of the biomarker to help focus the researcher on a particular goal and provide a record for future researchers. Some examples are shown in Table 2. These uses were developed by an ECETOC Task Force (ECETOC 2005) and are intended to focus the researcher towards a particular goal in the use of biomarkers.

Step two rigorously defines the exposure question against which the biomarkers will be evaluated. The exposure question should be well defined and may be the most important and difficult step in the Framework. The question should simultaneously address all five of the following critical elements:

- the Agent, or the substance for which the biomarker is an indicator,
- the Population the biomarker will be used with,
- the Route(s) of Exposure the biomarker is expected to detect,
- the Duration of Exposure the biomarker is expected to detect, and
- the Exposure Metric, such as cumulative exposure, mean exposure, peak exposure, etc.

These elements are also part of the ECETOC Task Force guidance (ECETOC 2005).

Examples of well defined questions would be:

- What would be the most scientifically supportable biomarker for assessing 30-year trends in cumulative lead exposure integrating all routes of exposure in children aged one to five?
- What is the most scientifically supportable biomarker for assessing average exposure to inorganic arsenic in the general population from drinking water over a one-year period?
- What is the most scientifically supportable biomarker for assessing average dermal occupational exposure to atrazine in farm workers over a 24-hour time period?

Step three is to develop the list of competing biomarkers and matrices to answer the question framed in the first two steps. To minimize total effort, only reasonable and viable biomarker/matrix combinations should be entered. The biomarker/matrix pairs will be evaluated and compared against each other. Each biomarker/matrix pair is independent of the others, so different matrices may be evaluated in a single analysis.

Table 2. List of potential intended uses for a biomonitoring analysis.

Potential Intended Uses of Biomonitoring Analysis	Description
General Purpose	Generic assessment of biomarkers.
Trend Assessment	Evaluate changes in exposure over time in a given population as measured by the biomarker.
Developing reference ranges	Evaluate concentrations of biomarker in a reference population (e.g. 5^{th} – 95^{th} percentile).
Evaluating new chemicals	Evaluate prevalence of biomarkers of emerging chemicals in a defined population.
Developing an exposure assessment method (e.g. environmental, occupational, or community)	Determine what biomarker to measure to assess exposure to a given chemical. Considers how specific the biomarker is to the given chemical or the best matrix for finding the biomarker.
Evaluating pre- and post-event personal exposures	Evaluating any exposure assessment involving the collection of samples before an exposure event and following exposure. For example, pre- and post- occupational sampling.
Assessing Health Outcomes	Evaluating associations between biomarker and health outcome.
Identifying Target Population (i.e. those at greatest risk for exposure)	Evaluating vulnerable or susceptible populations (e.g. children, pregnant women, elderly).
Evaluating efficacy of intervention efforts.	Evaluate population exposures before and after public health or regulatory intervention. For example, removing lead from gasoline.
Confirming previous reports	Confirming previous findings and conclusions from earlier studies/reports.
Using biomarkers for reverse dosimetry	Using biomonitoring data to assess exposure and dose.

The fourth step is to rate each criterion as to its *relevance* in answering the question posed in the second step. The criteria listed in Table 1 cover a wide range of attributes of a biomarker's intrinsic properties and while not strictly quantifiable, the attributes can be evaluated using a ranking system. As a guide, the list is divided into three sections dealing with issues of exposure and toxicology, sampling issues, and issues related to analytical methodology. The list is thought to be complete for almost all applications, but can be augmented in specific situations. It is suggested that the criteria be specified before the evaluation process begins. In the rating process, if a criterion is irrelevant it would be rated as 'No' or a numerical score of zero (a 'No' response effectively removes a criterion from the analysis). A relevant criterion would be rated as 'Yes' or a score of five. If the researcher is unsure of the relevance they can assign a rating of 'Maybe' or a score of three. The numerical values 0, 3, and 5 were chosen based on experimentation and experience with the process; these specific values are not the only possible choices, but work well in practice. These scores are called the "relevance" scores. As an example, considering one of the previously mentioned exposure questions, "What is the most scientifically supportable biomarker for assessing average exposure to inorganic arsenic in the general population from drinking water over a one-year period?," the criterion "Is knowing the half-life of the biomarker in the body important?" would be an important criterion because the question relates to 'average exposure over a one-year period' and would get a relevance score of 'Yes' or 5.

For step five each biomarker/matrix pair specified in step three is assigned a score indicating how suitable the biomarker is for meeting each criterion chosen in step four in conjunction with the question developed in the second step. The *suitability* is specified as 'None,' 'Low,' 'Medium,' and 'High,' with numerical values 0, 1, 3, and 5, respectively. As before, the specific numerical values of 0, 1, 3, and 5 are not the only possible choices but work well in practice. These scores are called the "suitability" scores. If a criterion has been characterized as 'No' for relevance in step four it is not necessary to specify its suitability in this step. *However, it is important that for each criterion with either a "Yes" or "Maybe" response in step four, a suitability score is assigned for* ***every*** *biomarker in step five.* Each biomarker can have only one response per criterion. Lastly, the response for each biomarker is independent from the others for each criterion.

In a small number of criteria a negative response is desirable. For example, for the criterion, "What is the cost per sample?" – A "High" response is less desirable, while a "Low" response is more desirable. Therefore, in cases where the criterion describes a negative characteristic, then the scores for the criterion are reversed. The criteria that are evaluated in this way are: To what degree is there inter-individual variation in the biomarker; what is the cost per sample; what is the total study cost; and, to what degree can contamination result in the presence of the biomarker in the matrix?

The choice of suitability score may seem somewhat subjective, especially between what constitutes a low, medium, or high score. Researchers will have to reconcile the choice of suitability scores for themselves. In developing the scores, it is valuable to have input of two or more experts.

The sixth and final step involves the required calculations and interpretation. For each biomarker/matrix pair listed from step three, the criterion relevance score is multiplied by the associated biomarker suitability score and the results summed over all criteria. The final score for each biomarker/matrix pair is calculated using all of the criteria included in the analysis, regardless of which of the three categories, or sub-sections, the criteria are in. A criterion item will contribute an increase to the overall score for a biomarker when it is considered an important criterion item *and* the biomarker is considered suitable for that criterion. A key feature of the Framework is that selection of criteria is flexible and can be tailored to evaluate specific questions of exposure. An unimportant criterion item *or* an unsuitable biomarker will make little or no contribution to the overall score, respectively. The summed score over all criteria is the raw score for the biomarker. To normalize the score on a scale of 0 to 100, the raw score is divided by the maximum possible score a biomarker could be assigned. The maximum score is the sum of the criterion relevance scores multiplied by 5, where 5 is the maximum suitability score.

B.Framework tool

All components of the Framework have been incorporated into a single Excel spreadsheet modeling tool called the Biomonitoring Weight of Evidence Calculator (WoEC) which guides the researcher through applying the interpretive criteria and calculating the biomarkers' ranking scores. The WoEC is available for downloading at: http://www.hesiglobal.org/i4a/pages/index.cfm?pageid=3488. The WoEC contains additional information and expanded definitions. We encourage readers to access the WoEC and explore the tool for their own applications.

In its present form, the WoEC spreadsheet is comprised of two primary worksheets. The first is labeled the "Evaluation" worksheet where all of the data input takes place. It is also where all of the tabulation of the biomarker scores takes place. The second primary worksheet is the "Results" worksheet. This worksheet was added to simplify the interpretation of the results for the researcher. No data input occurs here.

As a researcher enters the relevance and suitability scores, the WoEC automatically calculates the "Ranking" scores for each biomarker in the analysis. As a guide, the WoEC provides an expanded definition for each

criterion. At the bottom of each section of criteria, the WoEC calculates a normalized sub-score that summarizes the ranking of the biomarkers for all of the relevant criteria in that section. At the bottom of the Evaluation worksheet, the normalized final score is provided. The final scores account for the ranking of the biomarkers for all relevant criteria. As stated above, the final score for each biomarker is based equally on all of the relevant criteria, regardless of which of the three categories the criteria appear in. Each final score is normalized on a scale from 0 to 100 percent and is based only on those criteria that were included in the analysis. The biomarker with the highest normalized final "ranking" score is the most scientifically supportable biomarker to address the specific study question.

A researcher should note that the ranking of biomarker/matrix combinations is only as robust as the criteria that are included in an analysis. For example, although it is possible to obtain a result using just a single criterion to rank the biomarkers, this is discouraged since it underutilizes the power of the WoEC to interpret multiple biomarkers for a particular question of exposure.

C.Hypothetical calculation

The calculations are best illustrated with a simple hypothetical example. Suppose that an analysis has four competing biomarkers; A, B, C, and D. Further, suppose that the first three criteria had the responses shown in the top half of Table 3 for the relevance scores and the suitability scores. The resultant numerical values assigned to the relevance scores and suitability scores for each of the three criterion are shown in the bottom half of Table 3. Once the researcher has entered all of the "relevance" scores and "suitability" scores, the WoEC multiplies the "relevance" score for a criterion by the "suitability" score for each biomarker, as shown in Table 4. Next, the normalized sub-scores are calculated for each biomarker. The maximum possible sub-score for any biomarker in this example is 40 = (5 + 0+3) x 5, because the "relevance" scores are 5, 0, and 3, and a "suitability" score for any biomarker can be a maximum of 5 per criterion. The normalized sub-scores are obtained by dividing the sub-score for each biomarker by 40, and multiplying the quotient by 100. The normalized sub-scores for this simplified hypothetical example are in Table 5. The normalized final scores are obtained similarly as shown in the above example, except that the calculation is carried through for all criteria in the analysis.

Results

Use of the WoEC tool is demonstrated below for a benzene example. There is considerable interest in monitoring benzene exposure in both the occupational and community environment because of its toxicological effects and widespread occurrence. As the purpose of this paper is to illustrate the use of the WoEC tool rather than conduct an exhaustive review of benzene biomonitoring science, we rely on existing literature reviews and compendiums on exposure biomarkers for this chemical (ACGIH 2001, ATSDR 2007, DFG 1999, Farmer et al. 2005, Weisel 2010).

Step one of using the WoEC is to enter the intended use of the biomarker analysis. Of the multiple choices available in the drop down list (see Table 2), we chose "developing an exposure assessment method," to illustrate the use of the WoEC tool.

For step two, which requires entering the exposure question the biomarkers will be evaluated against, we chose to address the question, "what is a good biomarker for occupational benzene exposure through all routes over an eight-hour work shift that exceeds a concentration time-weighted average (TWA) of 1 ppm?" Note that there is a drop-down list available, or the researchers can enter a question of their choosing. The "question" (in step two) should address five critical elements: (1) the Agent (benzene), (2) the Population (workers), (3) the Exposure Route (inhalation), (4) the Assessed Timeframe (the total duration assessed, in this case eight-hour), and (5) the Exposure Metric (e.g. cumulative, mean, peak, etc., in this case exceeding a TWA concentration of 1 ppm).

Step three involves entering the biomarkers to be analyzed and the sample matrix for each biomarker. There are a number of potential biomarkers of benzene exposure that have been developed and deployed. These biomarkers have different degrees of sensitivity and are impacted by exposures other than benzene. Apart from benzene

Table 3. Illustration showing how numerical scores are assigned to the researcher-chosen "relevance" and "suitability" scores.

		Biomarker "Suitability" Score			
Criteria	"Relevance" Score	A	B	C	D
1	Yes	None	Low	Medium	High
2	No	Not relevant	Not relevant	Not relevant	Not relevant
3	Maybe	None	Low	Medium	High
		Corresponding Assigned Numerical Scores			
1	5	0	1	3	5
2	0	Not relevant	Not relevant	Not relevant	Not relevant
3	3	0	1	3	5

Table 4. Calculation of ranking scores* for hypothetical example.

	Biomarker Ranking Score = ("Relevance" Score) x ("Suitability" Score)			
Criteria	A	B	C	D
1	0	5	15	25
2	Not relevant	Not relevant	Not relevant	Not relevant
3	0	3	9	15
Sub-score†	0	8	24	40

*See bottom half of Table 3 for "relevance" scores and corresponding "suitability" scores.
†Sub-score is the sum of the biomarker ranking scores for each biomarker (i.e. column).

Table 5. Calculation of normalized sub-scores for hypothetical example.

	Biomarker			
Ranking Score	A	B	C	D
Sub-score*	0	8	24	40
Normalized Sub-score	0	20	60	100

*See Table 4 for sub-scores.

itself in exhaled air, blood or urine, other biomarkers include metabolites of benzene, such as phenol, hydroquinone, catechol, 1,2,4-trihydroxybenzene, and an open ring product called trans-trans muconic acid (tt-MA). In addition, s-phenlymercapturic acid (s-PMA), an adduct of glutathione, has also been used as a biomarker in urine (ACGIH 2001, DFG 1999, Farmer et al. 2005).

We selected six biomarkers of benzene exposure that were either the most commonly used biomarkers and/or had the most recent widespread usage. To simplify the analysis, the sample matrix for the six biomarkers selected was urine, although obviously other matrices could have been included. The final biomarkers chosen for this paper were benzene, catechol, hydroquinone, phenol, s-PMA, and tt-MA.

In step four, the researcher responds to the individual criterion based upon the question of exposure posed in step two. Once a researcher has determined that a particular section is relevant to the analysis, step five involves ranking each biomarker for its ability to address the specific criterion of interest. For this case study, suitability scores were developed for each biomarker independently as recommended in the Methodology section. The final criteria selected, associated rankings, normalized sub-scores, and a column documenting the scientific rationale for Section I, Exposure/Toxicology, can be viewed by following this link to the benzene example on the HESI's website: http://www.hesiglobal.org/i4a/pages/index.cfm?pageid=3488 (ILSI-HESI 2010). Section I can be found on the "EVALUATION" worksheet beginning on row 25. For Section II, Sampling, the criteria, final rankings, final sub-scores, and rationale are shown by following the same link to the benzene example on the HESI's website. Section II can be found on the "EVALUATION" worksheet beginning on row 51. Unlike the Exposure/Toxicology section, all criteria were deemed relevant for assessing the six benzene biomarkers/metabolites being evaluated. Lastly, Section III, Analytic Methodology, the criteria, rankings, sub-scores, and rationale are shown on the benzene example on the HESI's website. Section III can be found on the "EVALUATION" worksheet beginning on row 64. In this case study, we deemed several exposure-related criteria to be unimportant, and these criteria were set to NO. For example, for the criteria related to transport medium and exposure route, the exposure question of interest (see step two) seeks all routes, so a particular medium and/or exposure route was not important. For the remaining criteria, the rationale behind the rankings is indicated in the example. For example, regarding knowledge of exposure source, non-occupational benzene exposure such as tobacco smoking may be important with regard to contributing to the sources of benzene metabolites in urine. In this case, a specific biomarker of tobacco smoke exposure (e.g. cotinine in urine) should be included as an additional biomarker.

The calculations are completed in the WoEC. It is difficult to directly statistically determine the variation in the Framework normalized score without having done a full measurement analysis; we may do this type of analysis in the future if there is a need for it. To estimate a reasonable limit on the variation of a single score we need to consider there are 2 sources of variation in the final score, one from the variation within a researcher (or group of researchers if they jointly fill-out one form), the other the variation from researcher to researcher. From statistical principles we know that the variation in the score for a researcher depends on the number of criteria that are rated and the distribution of scores for a criterion. The number of criteria can change (if a researcher indicates the criterion is not relevant the criterion is not included in the final score), and the distribution of scores for criteria is unknown. However, the final score is normalized to be between 0 and 100, and it seems reasonable to assume that a meaningful interval of uncertainty around the normalized score is approximately ± 5 points. This may be considered as a rough estimate of a 90% credible interval. Similarly, under some reasonable assumptions of independence we can conclude that if two scores from the same researcher differ by less than 7.5 points, they are not practically different. We have no information

on the variability from researcher to researcher, but we may again assume that if the scores differ by less than 7.5 points, they are not practically different.

Table 6 shows the resulting scores for the six biomarkers that were evaluated. The conclusion in this case would be that, based on the ratings of the researcher, benzene in urine is the top ranked biomarker to assess occupational benzene exposure through all routes over an eight-hour work shift that exceeds a concentration time-weighted average of 1 ppm. This biomarker selection concurs with that which has been demonstrated in the literature for occupational monitoring of benzene (Farmer et al. 2005, Lovreglio et al. 2010). However, two other biomarkers, sPMA and ttMA in urine, were also highly ranked and may also be considered. Since hydroquinone, catechol, and phenol in urine, had lower rank scores, they are not suitable.

Discussion/Conclusions

We have developed a weight-of-evidence based framework for determining the scientific appropriateness of a biomarker for specific questions related to exposure biomonitoring. The framework was designed to be applied to one specific exposure question at a time and to compare results in a weight-of-evidence assessment for different biomarkers to help the assessor determine the most scientifically supportable biomarker to address the specific exposure in question. Currently, the framework is accessed through a simple spreadsheet, making it widely accessible and user friendly.

The formulation for the Framework is from the Planner Model (Roy and Nicolich 1980) that uses a consumer's awareness and preference for a set of product brands to determine the probable market share of a particular product; the Framework uses relevance and suitability in place of awareness and preference. The Framework is similar in format to the Analytic Hierarchy Process (AHP), which is a structured technique designed to aid making complex choices among competing decisions (Bhushan and Rai 2004). The AHP and the Framework are similar in that they each develop a series of criteria that are scored for importance, and they each list competing alternatives which are scored for each criterion. They differ in that for the AHP, the choices among the alternative decisions are made based on pair-wise choices among the competing alternatives, while for the Framework, the alternatives are based on assigned relevance scores. The final score under the AHP is based on a forced choice statistical model, and under the Framework the process is made under a linear model.

The Framework shares some characteristics of modern decision analysis (Clemen 1996) in that they both rely on 'expert judgment' to develop a final score and they both have a clearly delineated list of items that forms the basis for the final score. In both cases the expert scores for the list of items is clear and open for discussion and negotiation among the experts. In the Framework, the values selected for the relevance and suitability scores are subjective (steps four and five). Each researcher will use his or her expertise and experience in determining a score, and there will likely be differences in the scores of different researchers. An advantage of the Framework is that the researchers' opinions are broken out in discrete areas and choices and are available for inspection and discussion by others. It is likely that a unified assessment can be developed from differing opinions by using the clearly delineated scorings of the Framework. Three strong advantages of using the WoEC in applying the Framework are that (1) it provides space for the researcher to insert the rationale for their choices to facilitate later discussions, (2) it provides an item analysis by criterion to determine which criteria are critical in determining the researchers final score, and (3) it provides transparency and reproducibility in terms of criteria used and the scientific basis for the rankings, which allow for peer review and scientific consensus.

As an example of how the framework can promote peer review and scientific consensus, the authors formed two groups and each conducted an independent assessment for the benzene example question. The results from one of these assessments suggested that benzene in urine was the highest ranked biomarker, while the other independent assessment (not shown) suggested that sPMA in urine was the highest ranked. After discussion among the authors, and inspection of the suitability scores, it was revealed that the difference in the biomarkers depended largely on the timing of sample collection after exposure, with benzene in urine being most appropriate immediately after exposure occurred, but for longer intervals this biomarker suffers from volatility issues. Thus, the WoEC provided a transparent mechanism to facilitate scientific discussion and resolution.

The Framework provides for the first time a method to transparently document the rationale behind selecting, from among a set of alternatives, the most scientifically supportable exposure biomarker to address a specific biomonitoring question. There may be debate over the specific criteria and weighting factors included by an assessor using the WoEC; however, the criteria, weighting, and scoring are completely transparent, allowing other assessors the ability to reproduce and modify the assessment. Independent input to define the scores promotes scientific consensus and strengthens the assessment. Scores from several researchers independently completing the Framework for a mutually agreed upon question, set of biomarkers, and set of criteria can form the basis for a reasonable selection of a biomarker.

Table 6. Normalized final-scores for hypothetical example.

Ranked Order	Biomarker/Matrix	Score
1	Benzene / Urine	86
2	sPMA / Urine	81
3	ttMA / Urine	79
4	Hydroquinone / Urine	68
5	Catechol / Urine	58
6	Phenol / Urine	57

If there are disagreements among the researchers' final scores, the tabulations of the Framework show where the disagreements occur and form the basis for discussion to achieve a harmonized conclusion. The Framework also provides a common platform and enables the consistent application of criteria for facilitating a weight-of-evidence assessment and discussion among scientists regarding the selection of biomarkers of exposure.

As with any tool, application and use over time will allow refinement of the methodology. The current application of the Framework is focused on biomarkers of exposure, but a similar application can be made for biomarkers of effect or susceptibility.

Declaration of interest

The findings and conclusions in this paper are those of the authors and do not necessarily represent the official position of the Agency for Toxic Substances and Disease Registry. This publication stems from a subgroup of the HESI Integration of Biomonitoring Exposure Data into the Risk Assessment Process Technical Committee, whose work is funded through ILSI-HESI.

References

ATSDR (Agency for Toxic Substances and Disease Registry). (2007). Toxicological profile for Benzene. Atlanta, GA: U.S. Department of Health and Human Services, Public Health Service. Available: http://www.atsdr.cdc.gov/toxprofiles/tp3.html. Accessed 17 March 2010.

Albertini R, Bird M, Doerrer N, Needham L, Robison S, Sheldon L, et al. (2006). The use of biomonitoring data in exposure and human health risk assessments. *Environ Health Perspect* 114(11):1755–1762.

ACGIH (American Conference of Governmental Industrial Hygienists). (2001). TLVs and BEIs: Based on the documentation of the threshold limit values for chemical substances and physical agents and biological exposure indices. American Conference of Governmental Hygienists, Cincinnati, OH.

Bhushan N, Rai K. (2004). *Strategic Decision Making: Applying the Analytic Hierarchy Process*. New York: Springer Verlag.

Bird MG. (2008). Biomonitoring. In: Greim H and Snyder R, eds. *Toxicology and Risk Assessment: A Comprehensive Introduction*. Chichester, England: John Wiley & Sons.

Clemen R. (1996). *Making Hard Decisions: An Introduction to Decision Analysis*. 2nd ed. Belmont, CA: Duxbury Press.

DFG. (1999). List of MAK and BAT values. Commission for the Investigation of Health Hazards of Chemical Compounds in the Work area. Report No. 35 VCH. Weinheim, FRG: Deutsche Forschungsgemeinshaft.

Doerrer NG. (2007). Integration of human biomonitoring exposure data into risk assessment: HESI initiatives and perspectives. *Int J Hyg Environ Health* 210:247–251.

ECETOC. (2005). Guidance for the Interpretation of Biomonitoring Data. European Centre for Ecotoxicology and Toxicology of Chemicals, Document No. 44. Brussels, Belgium: European Centre for Ecotoxicology and Toxicology of Chemicals.

Farmer PB, Kaur B, Roach J, Levy L, Consonni D, Bertazzi PA, et al. (2005). The use of S-phenylmercapturic acid as a biomarker in molecular epidemiology studies of benzene. *Chem Biol Interact* 153–154:97–102.

ILSI-HESI (International Life Sciences Institute Health and Environmental Science Institute). 2008. Technical Committee on Integration of Biomonitoring Exposure Data into the Risk Assessment Process. Available: http://www.hesiglobal.org/NR/rdonlyres/9B242C4B-3819-4179-8B1C-6BA1C6EE2565/0/IntegrationofBiomonitoringExposureData.pdf. Accessed 17 June 2008.

ILSI-HESI (International Life Sciences Institute Health and Environmental Sciences Institute). 2010. Integration of Biomonitoring Exposure Data into the Risk Assessment Process Technical Committee: Biomarker of Exposure Interpretive Framework Work Group. Available: http://www.hesiglobal.org/i4a/pages/index.cfm?pageid=3488. Accessed 26 March 2010.

Lovreglio P, Barbieri A, Carrieri M, Sabatini L, Fracasso ME, Doria D, et al. (2010). Validity of new biomarkers of internal dose for use in the biological monitoring of occupational and environmental exposure to low concentrations of benzene and toluene. *Int Arch Occup Environ Health* 83:341–356.

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

Roy R, Nicolich M. (1980). Planner: A market positioning model. *Journal of Advertising Research* 20:61–66.

Weisel CP. (2010). Benzene: An overview of monitoring methods and their findings. *Chem Biol Interact* 184(1-2):58–66.

WHO (World Health Organization). (2001). Biomarkers in Risk Assessment: Validity and Validation. Environmental Health Criteria 222. Geneva: World Health Organization.