Regulatory Toxicology and Pharmacology 61 (2011) 185–191

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

Hypothesis-driven weight of evidence framework for evaluating data within the US EPA's Endocrine Disruptor Screening Program

Christopher J. Borgert [a,b,*], Ellen M. Mihaich [c], Lisa S. Ortego [d], Karin S. Bentley [e], Catherine M. Holmes [f], Steven L. Levine [g], Richard A. Becker [h]

[a] Applied Pharmacology & Toxicology, Inc., Gainesville, FL, USA
[b] C.E.H.T, University of Florida, Dept. Physiological Sciences, Gainesville, FL, USA
[c] Environmental and Regulatory Resources (ER^2), Durham, NC, USA
[d] Bayer CropScience, Research Triangle Park, NC, USA
[e] DuPont Crop Protection, Newark, DE, USA
[f] BASF Corporation, Research Triangle Park, NC, USA
[g] Monsanto Company, St. Louis, MO, USA
[h] American Chemistry Council, Washington, DC, USA

ARTICLE INFO

Article history:
Received 23 March 2011
Available online 23 July 2011

Keywords:
Weight of evidence
Endocrine Screening Program
Endocrine disruption
Regulatory framework

ABSTRACT

"Weight of Evidence" (WoE) approaches are often used to critically examine, prioritize, and integrate results from different types of studies to reach general conclusions. For assessing hormonally active agents, WoE evaluations are necessary to assess screening assays that identify potential interactions with components of the endocrine system, long-term reproductive and developmental toxicity tests that define adverse effects, mode of action studies aimed at identifying toxicological pathways underlying adverse effects, and toxicity, exposure and pharmacokinetic data to characterize potential risks. We describe a hypothesis-driven WoE approach for hormonally active agents and illustrate the approach by constructing hypotheses for testing the premise that a substance interacts as an agonist or antagonist with components of estrogen, androgen, or thyroid pathways or with components of the aromatase or steroidogenic enzyme systems for evaluating data within the US EPA's Endocrine Disruptor Screening Program. Published recommendations are used to evaluate data validity for testing each hypothesis and quantitative weightings are proposed to reflect two data parameters. Relevance weightings should be derived for each endpoint to reflect the degree to which it probes each specific hypothesis. Response weightings should be derived based on assay results from the test substance compared to the range of responses produced in the assay by the appropriate prototype hormone and positive and negative controls. Overall WoE scores should be derived based on response and relevance weightings and a WoE narrative developed to clearly describe the final determinations.

1. Introduction

On November 4, 2010, the US EPA released its draft "Weight-of-Evidence Guidance Document: Evaluating Results of EDSP Tier 1 Screening to Identify Candidate Chemicals for Tier 2 Testing" (US EPA, 2010). The Agency stated in its guidance that it would use WoE to determine whether a chemical has the potential to interact with the estrogen, androgen, or thyroid hormone components of the endocrine system. EPA stated that the intent of the document was "...to provide a transparent scientific approach for broadly evaluating Tier 1 screening data to determine if additional Tier 2 testing is necessary." EPA asserted its draft Guidance document provided a clear statement of how EPA intended to evaluate Tier 1 data so that the Agency's methodology would be transparent to all stakeholders.

The draft EPA WoE Guidance offers only some general considerations and principles related to making WoE determinations within the Endocrine Disruptor Screening and Testing Program (EDSP), and this may be viewed by some as providing a desired degree of flexibility for accommodating expert judgments within the effluvium of regulatory analyses and decision-making under uncertainty. However, the draft Guidance falls well short in describing how a WoE approach for the EDSP will be structured, how data will be evaluated for use in WoE, how the endpoints

* Corresponding author at: Applied Pharmacology & Toxicology, Inc., 2250 NW 24th Avenue, Gainesville, FL 32605, USA. Fax: +1 352 335 8242.
E-mail addresses: cjborgert@apt-pharmatox.com (C.J. Borgert), emihaich@nc.rr.com (E.M. Mihaich), lisa.ortego@bayer.com (L.S. Ortego), karin.s.bentley-1@usa.dupont.com (K.S. Bentley), catherine.holmes@basf.com (C.M. Holmes), steven.l.levine@monsanto.com (S.L. Levine), Rick_Becker@americanchemistry.com (R.A. Becker).

0273-2300/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.yrtph.2011.07.007

measured in the Agency's Tier 1 endocrine screening battery (ESB) will be weighted, or even how a weighing mechanism should be developed. A direct, transparent and objective methodology is still needed that will provide for consistency and credibility of WoE determinations made on the basis of EDSP data. A transparent and objective WoE methodology is especially necessary for the EDSP given the EPA's (and industry's) lack of experience conducting the ESB, the broad scope of the program, the significant impact inaccurate assessment could have on society and the regulated industry, and the excessive numbers of laboratory animals and costs required for Tier 2 testing.

The EDSP consists of two distinct tiers. Tier 1 is intended to determine whether a substance may interact with the endocrine system. Tier 1 consists only of screening assays, which are not sufficient alone to determine whether substances may have adverse health effects or to determine mode of action. Negative Tier 1 results would be adequate to determine that a substance is unlikely to have an effect on the estrogen, androgen or thyroid hormone systems or aromatase and steroidogenic enzymes. Positive Tier 1 results would indicate that the substance should be prioritized for Tier 2 testing. Tier 2, which consists of more apical assays, is intended to determine whether a substance may cause adverse effects, including those potentially mediated by the endocrine system, and evaluate the dose response associated with such effects. Tier 2 testing is more definitive than Tier 1 screening and negative Tier 2 results should supersede positive Tier 1 results (US EPA, 1998).

It is clear that screening assays provide qualitatively different information than definitive Tier 2 tests, and the results from these dissimilar assays should be used in a manner that is consistent with the scientific basis and purpose of each. The framework for conducting WoE evaluations for hormonally active agents proposed here is meant to operate within EPA's two-tiered EDSP and is intended to assist analysts in making the appropriate distinctions. Given the structure of EPA's EDSP, five separate WoE evaluations will be needed to assess EDSP data and to make the following determinations:

[a] determining from the Tier 1 ESB and other scientifically relevant information (OSRI) whether a substance exhibits the potential for interaction with androgen, estrogen, or thyroid pathways or aromatase and steroidogenic enzymes in vivo;
[b] determining from the Tier 1 ESB, OSRI and other information whether the substance should be further evaluated for endocrine activity in Tier 2 toxicity tests;
[c] determining from the results of Tier 2 toxicity tests whether a substance exhibits adverse effects potentially mediated by androgen, estrogen, or thyroid pathways;
[d] determining from Tier 1 ESB, OSRI, Tier 2 toxicity tests, and as necessary, additional mode-of-action experiments, whether the adverse effects observed in Tier 2 toxicity tests are a consequence of endocrine activity, and;
[e] determining whether endocrine-mediated adverse effects on humans or wildlife are possible at environmentally relevant exposure levels.

The framework for conducting WoE evaluations described here is applicable to all five of these separate determinations. This publication describes the elements of the framework, including its relationship to other published WoE approaches for endocrine active substances, the overarching scientific principles that govern data evaluation within the framework, and the two primary weighting types used to evaluate data for each WoE determination. This publication does not, however, describe the operational and technical details necessary to carry out the five individual WoE determinations. Subsequent publications will provide those. Instead, this paper focuses on the principles and processes for weighting data and illustrates how this is to be done for Tier 1 ESB data, i.e., for WoE determination [a] above.

Before delving further into the background literature and scientific principles governing the proposed framework, it is imperative to define terminology clearly so that the WoE framework can be considered in its proper context. Weed (2005) has noted that the term "weight of evidence" is used frequently in the scientific literature without being defined. According to Weed, the term is used in three categorically distinct ways: (1) metaphorical, (2) methodological, and (3) theoretical. As used in the framework proposed here, the term "weight of evidence" is both theoretical in that it labels the overall process, as well as methodological in that it describes specific methods and qualitative principles governing the use of the proposed process. In subsequent publications, various quantitative procedures will be described that might be used to weight data from the various types of studies relevant for evaluating potential endocrine activity and endocrine-mediated toxicity. Importantly, the framework proposed here incorporates step-by-step documentation and transparency of the decision process, which have been identified as elements that enhance scientific credibility (Borgert, 2007a,b; Schreider et al., 2010).

The proposed WoE approach can be summarized according to the following seven steps, the justification and background (Section 2), scientific principles (Section 3), operational details (Sections 4 and 5), and implications (Section 6) of which are explained further in this paper and in the tabular summaries available as Supplementary material:

1. define specific hypotheses to be evaluated;
2. systematically search, review and select data relevant to each hypothesis;
3. evaluate the primary validity and reliability of each study selected, and for WoE evaluations involving causality (e.g., [c] and [d] above), determine whether the data are derived from counterfactually designed studies;
4. develop quantitative or rank ordered relevance weightings (W_{REL}) for each type of assay or endpoint with respect to its sensitivity and specificity for testing the hypothesis;
5. develop quantitative response weightings (W_{RES}) based on results for the test substance compared to positive and negative controls in each assay or endpoint;
6. combine relevance (W_{REL}) and response (W_{RES}) weightings according to a pre-defined algorithm to produce an overall WoE score;
7. develop an overall WoE determination as to whether each hypothesis is supported or rejected, and how strongly, based on the overall WoE scores.

2. Background and justification

Several organizations have developed frameworks and discussed principles important for conducting WoE evaluations (Balls et al., 2006; Bars et al., 2011; Boobis et al., 2006, 2008; Damstra et al., 2002; ECETOC, 2009; Gray et al., 2001; Menzie et al., 1996) and independent investigators have published WoE frameworks and evaluations of endocrine active substances (e.g., Calabrese et al., 1997; Goodman et al., 2006, 2009; Martin et al., 2007; Rhomberg, 1998, 2008; Rhomberg and Goodman, 2008). It is beyond our scope to summarize each of these frameworks and publications, but a general overview is provided in the overview of weight of evidence frameworks in Supplementary material, which is helpful for understanding overarching issues related to developing WoE frameworks and is essential for understanding our proposed framework in the context of this previous work.

None of the previous frameworks, however, were specifically formulated for making WoE decisions within the US EPA's Tier 1 EDSP, and do not provide 'off-the-shelf' readiness fit for this specific purpose. In particular, those frameworks do not address the question of deciding whether a substance should be subjected to Tier 2 testing based on the results of Tier 1 ESB data and integrated with existing OSRI. Now that the EDSP 890 Series Test Guidelines (US EPA, 2009) have been issued, it is possible to develop a more complete and specific WoE framework to address the particular needs of the EDSP program. It was therefore determined that a new, thorough WoE approach should be developed, specifically tailored for use in the context of the EDSP. This new approach should incorporate and build upon strengths of the prior work done in this area, taking care to avoid pitfalls and fill critical deficiencies noted by analysts who have surveyed WoE methodologies (Krimsky, 2005; Schreider et al., 2010; Weed, 2005). In contrast to previously mentioned frameworks, this new approach should clearly define its scientific foundation, should provide a mechanism for actually 'weighing' evidence, and should clearly describe the derivation of that weighting mechanism.

The approach proposed here seeks to leverage the considerable strengths of published WoE frameworks and evaluations. Chief among these is a focus on specific hypotheses to be evaluated by the WoE methodology, as advocated by Rhomberg and colleagues (Goodman et al., 2006, 2009; Rhomberg, 1998; Rhomberg and Goodman, 2008). Unambiguous hypotheses enhance the clarity with which assays and endpoints can be assigned as relevant for specific evaluations. Furthermore, clear hypotheses should allow a more focused weighting of the various assays and endpoints on the basis of empirical data rather than the application of expert judgment alone. The hypothesis-driven basis of the proposed approach will ideally allow assignment of a quantitative relevance weighting for each endpoint and each assay in the Tier 1 ESB. At the least, it should allow a semi-quantitative rank ordering of the assays and endpoints based on empirical observations.

3. Overarching scientific principles

In developing a WoE framework, a number of overarching scientific principles must be considered. First, relevant, testable hypotheses should be developed. Data relevant to each hypothesis should be gathered from mandated guideline studies or from the literature according to clearly stated methods and criteria for search and selection, similar to rules used to develop systematic reviews (Farquhar and Vail, 2006; Gronseth, 2004; McQueen, 2001; Oosterhuis et al., 2004; Ricci et al., 2006; Smyth, 2000; Zaza et al., 2000). Data should then be evaluated in terms of minimal epistemic status, reliability, and probative nature of the study design for evidence of causation (Borgert and Gori, in preparation; Gori, 1999, 2001, 2002, 2009a, 2009b, 2010; Klimisch et al., 1997; Schneider et al., 2009; Subcommittee on Energy and Environment, 2010; Subcommittee on Health, 2010), which we designate primary, secondary, and tertiary validity, respectively. A detailed discussion of these fundamental scientific principles and how they relate to the WoE framework proposed here for the ESDP is provided in Building the Weight of Evidence Framework on Scientific Principles found in Supplementary material. The Supplementary material is essential for understanding the proposed WoE framework, and the approach cannot be evaluated or applied without it. It is important to clarify that the purpose of these principles is not to exclude data on the basis of low primary, secondary, or tertiary validity, but that this evaluation be reflected in the overall relevance weighting W_{REL} of the endpoint or data.

4. WoE determination [a]: potential activity in Tier 1 screening battery

In order to illustrate the proposed approach, details of the framework for WoE evaluation [a] are described. As discussed above, the first step in developing a WoE framework is to develop relevant hypotheses. Using the US EPA's Tier 1 ESB as a working example, eight discrete hypotheses are proposed for conducting a WoE evaluation for determination [a] – determining whether a substance exhibits the potential to interact with androgen, estrogen, or thyroid pathways in vivo, as follows:

[a] The chemical exhibits the potential to:
[a]-1 interact as an agonist with components of estrogen pathways.
[a]-2 interact as an antagonist with components of estrogen pathways.
[a]-3 interact as an agonist with components of androgen pathways.
[a]-4 interact as an antagonist with components of androgen pathways.
[a]-5 interact as an agonist with components of thyroid pathways.
[a]-6 interact as an antagonist with components of thyroid pathways.
[a]-7 interact with components of aromatase enzyme system.
[a]-8 interact with components of steroidogenesis enzyme system.

Evaluating each hypothesis requires a clear understanding of the results for each assay based on prototypical active (positive control) and inactive (negative control) chemicals in each of the EDSP Tier 1 ESB assays. This understanding should include an evaluation of the validity of each endpoint measured in the assay as described in Section 3 above. Ideally, the analyst's understanding would also include the positive and negative predictive value of each endpoint in each assay – i.e., its sensitivity and specificity – for predicting adverse endocrine-mediated effects in Tier 2 tests. Both aspects should be accounted for by a quantitative weighting of each endpoint for use in the WoE evaluation. This will provide a "relevance weighting" for each endpoint and assay, for each hypothesis, denoted by W_{REL}. An alternative to this quantitative assignment of W_{REL} values would be to assign a rank ordering of relevance for each endpoint for each hypothesis.

It is important to understand that W_{REL} values (or rank orderings) need to be assigned for each endpoint that is relevant for each of the eight hypotheses [a]-1 through [a]-8. These W_{REL} values would be expected to differ for each hypothesis, but not all endpoints have relevance for each hypothesis. In other words, some endpoints will have a W_{REL} value of zero for some hypotheses, irrespective of the strength and validity of that endpoint for its intended purpose. For example, a zero W_{REL} value might be assigned to the uterotrophic response for evaluating hypothesis [a]-6, that a chemical has potential to interact as an antagonist with components of thyroid pathways, even though it may have a very high W_{REL} value for evaluating hypothesis [a]-1 (potential to interact as an agonist with components of the estrogen pathway). It is also important to understand that different W_{REL} values (or rank orderings) may need to be assigned for different WoE determinations [a] through [e]. In other words, the W_{REL} value for the uterotrophic response may be quite different for evaluating hypothesis [a]-1 (estrogen agonist pathway) than for hypotheses related to WoE determination [c], whether a substance exhibits adverse effects potentially mediated by androgen, estrogen, or thyroid pathways.

Because the validation programs for endocrine screening assays included in the US EPA's Tier 1 ESB did not include a formal

Table 1
Male Pubertal Assay Example – Hypothesis [a]-6; Test agent acts as an antagonist with components of thyroid pathways in vivo.

Tier 1 Assay Pubertal Male	Tier 1 Assay Endpoints	Prototypical Response* of Thyroid Antagonist	Response* of Negative Control	ESB Response* of Test Agent W_{RES}	OSRI Response* of Test Agent W_{RES}	Relevance Weighting** for Endpoint W_{REL}
	Growth (daily body weight) Age and weight at preputial separation Seminal ves. + coag. gland weight Ventral prostate weight Dorsolateral prostate weight Levator ani + bulbocavernosus Muscle complex weight Epididymis weight Testes weight Thyroid weight Adrenal weight Pituitary weight Blood chemistry, standard panel Hormone levels Testes histopathology Thyroid histopathology Epididymides histopathology	Obtain data from EPA's method validation studies	Obtain data from EPA's method validation studies	Obtain data from the study run in response to EDSP test order	Obtain data from scientific literature & evaluate data validity & reliability	Obtain W_{REL} from consensus workshop

* To include consideration of dose response, statistical significance and biological significance.
** WREL values for the various endpoints would be expected to differ depending upon the hypothesis under evaluation. Some endpoints might have no relevance for a particular hypothesis.

assessment of positive and negative predictive value, that information is unlikely to be available until the results of screening assays for the first EDSP pesticide chemicals placed under test order are completed, compiled, and evaluated. In lieu of those data and evaluations, an interim approach for assigning W_{REL} values would be to assign the weighting based on an evaluation of the relative validity of the endpoints and any data available from relevant and reliable published literature and from regulatory GLP studies. This process and the data used should be fully described in the WoE evaluation, in accordance with principles described here for evaluating primary, secondary and tertiary validity. Individual analysts could assign W_{REL} values on a case-by-case basis; however, greater consistency would be achieved by holding expert consensus workshops to establish W_{REL} values. The consensus workshops could decide whether W_{REL} values or rank orderings are more useful and practical, and similar panels could update these as more complete information emerges from the first round of EDSP screening.

The actual results of each assay for each substance evaluated in the ESB can be arrayed alongside the "expected results" of prototypical active and inactive agents. This comparison would seem to be most transparent if a quantitative rating is also given to the response produced by the test article. The weighting should reflect the fact that the endpoints measured are typically continuous variables and not ordered responses. Thus, we propose that the response of the test article be weighted according to a scale that assigns the prototypical active agent, or positive control, the highest positive weighting for each endpoint and the negative control the lowest. This addresses the ECETOC recommendation to consider potency in WoE determinations regarding all aspects of endocrine activity, and addresses Krimsky's criticism that WoE methodologies often involve transforming continuous data to dichotomous or triadic variables via black box judgments. This proposed "response weighting" is denoted by W_{RES}, for the test article at the specific endpoint or assay.

The WoE evaluation for Tier 1 ESB would proceed for each hypothesis by completing a tabular summary for each hypothesis [a]-1 through [a]-8. The first tabular summary involves input of data from the endpoints measured in each Tier 1 ESB, as well as data obtained from published literature or other sources (e.g., OSRI). To illustrate the concept, Table 1 shows a portion of this first tabular summary listing endpoints from the male pubertal assay to be evaluated for Hypothesis [a]-6, interaction as a thyroid hormone antagonist. This process would be conducted for each hypothesis [a]-1 through [a]-8, so that a complete tabular summary would be repeated for each hypothesis (see Supplementary material, Table A). The second tabular summary (see Supplementary material, Table B) compiles the results of the first tabular summary by deriving an overall weighting for each assay for each hypothesis; Supplementary material Table B provides an example using Hypothesis [a]-6. Table B would also be repeated for each hypothesis, resulting in 8 iterations, one for each hypothesis [a]-1 through [a]-8. The third tabular summary (Table 2) compiles the results of all eight iterations of Table B. A WoE narrative should accompany Table 2 to describe the overall determination as to whether the actual data from the test article sustains or refutes each hypothesis [a]-1 through [a]-8, supported by an explanation and rationale that includes comparison of WoE scores for known positive and negative substances. All mitigating circumstances or conditions should be described, such as equivocal results in some assays. It is recommended that the analyst or expert panel derive a standard lexicon for stating the results of the determinations, similar to the recommendations of IPCS (Boobis et al., 2006, 2008). For example, a standardized determination might be:

> *"Evidence is considered sufficient/insufficient/equivocal to indicate the test article, under conditions of the experiments, has the potential to interact with one or more components of the Estrogen pathway system in an agonistic manner." Rationale: . . ."*

5. WoE determinations [b] through [e]

WoE determinations [b] through [e] are not developed in detail in this report, but will comprise subsequent publications. In general, the basic components of the proposed WoE framework should

Table 2
WoE Determination [a]: determining from the Tier 1 ESB and OSRI whether a substance exhibits the potential for interaction with androgen, estrogen, or thyroid pathways in vivo

Hypotheses	Results of Hypothesis Based WoE Evaluation $\sum (\sum(W_{REL} \times W_{RES})$
1. Hypothesis: chemical interacts as an agonist with components of estrogen pathways 2. Hypothesis: chemical interacts as an antagonist with components of estrogen pathways 3. Hypothesis: chemical interacts as an agonist with components of androgen pathways 4. Hypothesis: chemical interacts as an antagonist with components of androgen pathways 5. Hypothesis: chemical interacts as an agonist with components of thyroid pathways 6. Hypothesis: chemical interacts as an antagonist with components of thyroid pathways 7. Hypothesis: chemical interacts with components of the aromatase enzyme system 8. Hypothesis: chemical interacts with components of the steroidogenesis enzyme system	Obtain $\sum(W_{REL} \times W_{RES})$ values for each assay from Table B (Supplementary materials); Combine scores for all assays relevant to each hypothesis.

A WoE narrative should accompany Table 2 to describe the overall determination as to whether the actual data from the test article supports or refutes each hypothesis [a]-1 through [a]-8. Each determination should be supported by an explanation and rationale that includes comparison of WoE scores for known positive and negative substances. All mitigating circumstances or conditions should be described, such as equivocal results in some assays.

be applied to each of the five WoE determinations. For WoE determination [b] – deciding whether Tier 2 testing is warranted – a consideration should be made of OSRI and other toxicity information to determine whether the potency for a potential endocrine activity could manifest actual toxicity that would be relevant for human health or for human or ecological risk assessment. For WoE determinations [c] through [e], the proposed method should be used for weighting relevance and response as a first step before proceeding to the flow charts presented in the ECETOC guidance (Balls et al., 2006; ECETOC, 2009). For each step stated or implied by the ECETOC framework, explicit hypotheses should be developed, and the principles and criteria for reliability, validity, and transparency outlined herein should be applied in evaluating the data used in these determinations.

The order of WoE determinations proposed here is not rigid. For protecting public health, it may be more important to determine whether the potential endocrine-mediated effects identified in Tier 2 testing (WoE determination [c]) are possible at levels of exposure occurring in humans (WoE determination [e]) than to determine unequivocally that the adverse effects are indeed produced by an endocrine mechanism (WoE determination [d]). The salience of this point owes, in part, to the difficulty of fully establishing the mechanism of action for any particular effect, especially for chemicals that produce a variety of biological effects within the same dose range. For systemic toxicants, it may be challenging to differentiate toxic effects at target organs mediated by non-endocrine modes of action from weak endocrine activity. Furthermore, the types of adverse effect endpoints evaluated in Tier 2 are not specific to a particular endocrine pathway and may also occur as a result of toxicity unrelated to any endocrine activity. Hence, these may or may not be causally associated with Tier 1 results.

6. Discussion

Scientific conclusions in the health and environmental fields are rarely made on the basis of a single evidentiary modality, thus WoE approaches are needed to draw evidence from a variety of different modalities with the expectation of convergence upon a coherent and consistent conclusion (Krimsky, 2005). Krimsky explains that because convergence is often elusive and data are often contradictory, WoE approaches designed to "test" or evaluate convergence typically rely on expert judgments to declare which lines of evidence are stronger and thus, which evidence should be weighted most heavily. Although Krimsky's criticism of such judgments seems sound, the solutions proposed here are not to improve expert judgments but to enhance the role of hypothesis generation and testing in WoE evaluations.

Although it is impossible to completely remove expert judgment from the scientific process, the scientific method itself calls for removal of the scientist or analyst as far as possible from the generation and interpretation of data. To this end, the WoE methodology proposed here is intended to render as much of the decision-making process as possible data-dependent. In order to do this, the framework relocates judgmental aspects that are typically made in analysis and interpretation to the process of methodological formulation, and this methodology has in turn been tied as closely as possible to inference based on data from specific assays and endpoints. This should not be interpreted as suggesting box checking over scientific judgment, but rather as elevating well-formulated inference based on objective data above so-called "expert judgments." This promotes standardization of data review and evaluation approaches, encourages use of standard procedures for evidence identification, focuses analysis based on an understanding of mode of action, and standardizes the approach for conducting the weight-of-evidence evaluation. These objectives and design features are indistinguishable from the recent recommendations of the National Research Council for improving EPA risk assessments (NRC, 2011).

At the same time, it must be acknowledged that experimental selection and design involves scientific judgment, and that even this process entails a subjective component. This is addressed with a formal statement about, and criteria by which to judge, the validity of the data brought to bear in the WoE evaluation. Although these procedures cannot completely remove expert judgment from the WoE process, they afford a means of dealing transparently with these issues as objectively as possible.

Some may see this WoE framework as one pole of the spectrum of approaches a regulatory agency may use to make decisions, with the opposite pole being extreme interpretation of the precautionary principle. It is acknowledged that the proposed approach may appear tedious, resource intensive, and that equivalent requisite data are unlikely to be available for many of the WoE

determinations, rendering any assessment based on those data imperfect. It would seem, however, that the numerous imperfections that may arise from disparate datasets are at least identifiable and therefore transparent, and pale in contrast to the numerous but hidden imperfections that are certain to arise by WoE processes wholly dependent on expert judgments.

Clearly, the level of detail, rigor, and quantification within a WoE evaluation process should be commensurate with the nature or impact of the decision being reached. In addition to risk, benefits, cost considerations, and societal concerns all play important roles in regulatory policies and decision-making. Regulatory decisions must consider economic costs of testing and regulation, the need for health protection and the potential costs of adverse health effects. But, as the Bipartisan Policy Center (BPC, 2009) panel concluded in their report entitled Improving the Use of Science in Regulatory Policy, "*science can inform some policy choices, but it can't determine them*," and that regulatory processes must be developed that "*explicitly differentiate, to the extent possible, between questions that involve scientific judgments and questions that involve judgments about economics, ethics and other matters of policy*." Moreover, the considerable resources necessary to conduct WoE evaluations by the method proposed here would seem well warranted for a program that could represent a burden to the US economy in excess of one hundred million dollars (Borgert et al., 2011). This proposed WoE method aligns with these BPC recommendations, as evidenced by its focus on increasing the scientific rigor and transparency of endocrine WoE analyses.

Despite making specific methodological recommendations about hypothesis formulation, evaluation of data validity, and development of quantitative weightings, no specific recommendations are made here regarding several likely eventualities of endocrine screening and testing. These will need to be addressed either by individual analysts, consensus workshops, or future publications. Although it is beyond the scope of this overall framework to address each possible eventuality, several should be mentioned for completeness.

For example, since the repetition of certain assays or similar types of assays is likely, both confirmatory and contradictory results are possible from endocrine screening and toxicology studies. While it would seem reasonable for confirmatory results to increase and contradictory results to decrease overall WoE weightings, it is not immediately apparent how a simple quantitative algorithm could address this consideration without subjecting WoE determinations to the potential influence of the replication number for certain assays. Similarly, the possibility exists for in vitro and in vivo results to differ and for epidemiological data to contradict data from experimental laboratory systems. While there would seem to be widespread agreement that, in general, in vivo assays should be given more credence than in vitro assays (US EPA, 2010), this generalization may not be appropriate in all instances and professional judgment will be required to properly weight evidence from different biological levels of organization. Furthermore, 'strength of evidence' may be an important concept for considering some of these issues. The ECETOC framework for combining animal and human data elevates data based on the quality of the study (Bars et al., 2011), which is appropriate in many contexts, but the relevance of the test system must also be considered. Regardless of whether or not analysts impose a hierarchy of data types used to resolve potential contradictions, it is important that the rationale for doing so (or not) be provided and if imposed, the methodology described in detail and applied consistently. In addition, no recommendation is given here as to whether a threshold of evidence is necessary for any particular WoE decision. This may be desirable and necessary, but the details for establishing such thresholds should be developed after W_{REL} values are established for the determination, hence, it would be premature to do in this general framework.

Finally, new methodologies should ultimately be tested to confirm accuracy and utility. However, since "weight of evidence" methodologies are developed specifically to fill a gap created by the lack of a definitive test, the only means of 'testing' them is against other judgmental processes. One way of testing the approach proposed here would be to use it to evaluate, in a blinded manner, data generated for chemicals with well-known endocrine properties that produce specific adverse effects, as well as data for chemicals that clearly lack such activity and toxicity. Use of the proposed methodology should confirm the known activities; if deviations are found, these could be used to refine the process itself or the individual W_{REL} scores.

In conclusion, the proposed approach builds upon recent advancements in the practice of WoE by incorporating the greatest strengths of recently published frameworks and methodologies. As such, it represents an advancement of previous work because it provides a clear statement of the foundation upon which interpretations are made and proposes specific criteria for evaluating the status of the data used in the WoE evaluations. These improvements are necessary to satisfy many legitimate criticisms delineated by Krimsky (2005) and Weed (2005) and to ensure that WoE evaluations conducted under the US EPA's EDSP are objective, transparent, consistent, and scientifically reliable. Toward that end, the convening of consensus workshops should occur as soon as possible to begin the critical work of developing objective and transparent relevance weightings (W_{REL}) for all of the Tier 1 ESB endpoints that will be used in making WoE determinations under the EDSP. Future publications are in development that will propose a starting point for these important deliberations regarding the WoE methodology for Tier 1 ESB endpoints.

Conflict of interest statement and funding disclosure

Some substances undergoing EDSP screening are compounds produced by Monsanto, Bayer, DuPont, and BASF. However, this manuscript applies to all substances impacted by EPA EDSP test orders, not merely those produced by the authors' employers. This manuscript has been reviewed in accordance with the peer- and administrative-review policies of the authors' organizations. The views expressed here are those of the authors and do not necessarily reflect the opinions and/or policies of their employers. There are no contractual relations or proprietary considerations that restrict dissemination of the research findings of the authors. C.J. Borgert, and E.M. Mihaich are independent scientists/consultants who received financial support for portions of this project from the Endocrine Policy Forum. Time spent by other co-authors was supported by their respective employers or a personal contribution.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.yrtph.2011.07.007.

References

Balls, M., Amcoff, P., Bremer, S., Casati, S., Coecke, S., Clotheir, R., et al., 2006. The principles of weight of evidence validation of test methods and testing strategies: the report and recommendations of ECVAM Workshop 58a. Alternatives to Laboratory Animals 34 (6), 603–620.

Bars, R., Broeckaert, F., Fegert, I., Gross, M., Hallmark, N., Kedwards, T., et al., 2011. Science based guidance for the assessment of endocrine disrupting properties of chemicals. Regulatory Toxicology and Pharmacology 59, 37–46.

Boobis, A.R., Cohen, S.M., Dellarco, V., McGregor, D., Meek, M.E., Vickers, C., et al., 2006. IPCS framework for analyzing the relevance of a cancer mode of action for humans. Critical Reviews in Toxicology 36 (10), 781–792.

Boobis, A.R., Doe, J.E., Heinrch-Hirsch, B., Meek, M.E., Munn, S., Ruchirawat, M., et al., 2008. IPCS framework for analyzing the relevance of a noncancer mode of action for humans. Critical Reviews in Toxicology 38 (2), 87–96.

Borgert, C.J., 2007a. Conflict of interest or contravention of science? Regulatory Toxicology and Pharmacology 48 (1), 4–5.

Borgert, C.J., 2007b. Conflict of interest: kill the messenger or follow the data? Environmental Science & Technology 41 (3), 665.

Borgert, C.J., Gori, G.B., in preparation. Demonstrating causation: essential concepts from pharmacology. Toxicology and Epidemiology.

Borgert, C.J., Mihaich, E.M., Quill, T.F., Marty, M.S., Levine, S.L., Becker, R.A., 2011. Evaluation of EPA's tier 1 endocrine screening battery and recommendations for improving the interpretation of screening results. Regulatory Toxicology & Pharmacology 59 (3), 397–411.

Bipartican Policy Center, 2009. Improving the use of Science in Regulatory Policy. <http://bipartisanpolicy.org/sites/default/files/BPC%20Science%20Report%20fnl.pdf> (accessed 16.06.11).

Calabrese, E.J., Baldwin, L.A., Kostecki, P.T., Potter, T.L., 1997. A toxicologically based weight-of-evidence methodology for the relative ranking of chemicals of endocrine disruption potential. Regulatory Toxicology and Pharmacology 26 (1), 36–40.

Damstra, T., Barlow, S., Bergman, A., Kavlock, R., Van der Kraak, G., 2002. Causal criteria for assessing endocrine disruptors: a proposed framework. In: Global Assessment of the State-of-the-science of Endocrine Disruptors. World Health Organisation, International Programme on Chemical Safety (Chapter 7)

ECETOC, 2009. Technical Document 106: Guidance on Identifying Endocrine Disrupting Effects. European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC), Brussels, Belgium.

Farquhar, C., Vail, A., 2006. Pitfalls in systematic reviews. Current Opinion in Obstetrics and Gynecology 18 (4), 433.

Goodman, J.E., McConnell, E.E., Sipes, I.G., Witorsch, R.J., Slayton, T.M., Yu, C.J., Lewis, A.S., Rhomberg, L.R., 2006. An updated weight of the evidence evaluation of reproductive and developmental effects of low doses of bisphenol A. Critical Reviews in Toxicology 36, 387–457.

Goodman, J.E., Witorsch, R.J., McConnell, E.E., Sipes, I.G., Slayton, T.M., Yu, C.J., Franz, A.M., Rhomberg, L.R., 2009. Weight-of-evidence evaluation of reproductive and developmental effects of low doses of bisphenol A. Critical Reviews in Toxicology 39, 1–75.

Gori, G.B., 1999. The EPA and the courts: inching toward a showdown. Regulatory Toxicology and Pharmacology 30, 167–168.

Gori, G.B., 2001. The costly illusion of regulating unknowable risks. Regulatory Toxicology and Pharmacology 34 (3), 205–212.

Gori, G.B., 2002. Considerations on guidelines of epidemiologic practice. Annals of Epidemiology 12 (2), 73–78.

Gori, G.B., 2009a. Conflict of interest and public policy. Regulatory Toxicology and Pharmacology 53 (3), 159–160.

Gori, G.B., 2009b. Scientific integrity. Regulatory Toxicology and Pharmacology 54 (3), 213.

Gori, G.B., 2010. Regulating unknown risk. Regulation 33 (1), 16–21.

Gray, G.M., Baskin, S.I., Charnley, G., Cohen, J.T., Gold, L.S., Kerkvliet, N.I., et al., 2001. The Annapolis Accords on the use of toxicology in risk assessment and decision-making: an Annapolis Center workshop report. Toxicology Mechanisms and Methods 11, 225–231.

Gronseth, G.S., 2004. From evidence to action. Neurorx 1 (3), 331–340.

Klimisch, H.J., Andreae, M., Tillmann, U., 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regulatory Toxicology and Pharmacology 25 (1), 1–5.

Krimsky, S., 2005. The weight of scientific evidence in policy and law. American Journal of Public Health 95 (Suppl. 1), S129–S136.

Martin, O.V., Lester, J.N., Voulvoulis, N., Boobis, A.R., 2007. Human health and endocrine disruption: a simple multicriteria framework for the qualitative assessment of end point specific risks in a context of scientific uncertainty. Toxicological Sciences 98 (2), 332–347.

McQueen, M.J., 2001. Overview of evidence-based medicine: challenges for evidence-based laboratory medicine. Clinical Chemistry 47 (8), 1536–1546.

Menzie, C., Henning, M.H., Cura, J., Finkelstein, K., Gentile, J., Maughan, J., et al., 1996. Special report of the Massachusetts weight-of-evidence workgroup A weight-of-evidence approach for evaluating ecological risks. Human and Ecological Risk Assessment: An International Journal 2 (2), 277–304.

NRC 2011. Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde, A Roadmap for Revision. National Research Council, National Academy Press, Washington, D.C (Chapter 7) <http://www.nap.edu/catalog/13142.html> (accessed 16.06.11).

Oosterhuis, W.P., Bruns, D.E., Watine, J., Sandberg, S., Horvath, A.R., 2004. Evidence-based guidelines in laboratory medicine: principles and methods. Clinical Chemistry 50 (5), 806–818.

Rhomberg, L.R., 1998. Risk characterization for human health effects from potential environmental endocrine disruptors. In: Dunaif, G.E., Olin, S.S., Scimeca, J.A., Thomas, J.A. (Eds.), Human Diet and Endocrine Modulation: Estrogenic and Androgenic Effects. ILSI Press.

Rhomberg, L.R., 2008. A Framework for Weight of Evidence: Application to Endocrine Effects and the Low-dose hypotheses. ISRTP Workshop: Conducting and Assessing the Results of Endocrine Screening. Bethesda, MD.

Rhomberg, L.R., Goodman, J.E., 2008. CERHR conclusions would have been strengthened by a more explicit weight-of-evidence analysis. Birth Defects Research Part B: Developmental and Reproductive Toxicology 83, 155–156.

Ricci, S., Celani, M.G., Righetti, E., 2006. Development of clinical guidelines: methodological and practical issues. Neurological Sciences 27 (Suppl. 3), S228–S230.

Schneider, K., Schwarz, M., Burkholder, I., Kopp-Schneider, A., Edler, L., Kinsner-Ovaskainen, A., et al., 2009. "Toxrtool", a new tool to assess the reliability of toxicological data. Toxicology Letters 189 (2), 138–144.

Schreider, J., Barrow, C., Birchfield, N., Dearfield, K., Devlin, D., Henry, S., et al., 2010. Enhancing the credibility of decisions based on scientific conclusions: transparency is imperative. Toxicological Sciences 116 (1), 5–7.

Smyth, R.L., 2000. Evidence-based medicine. Paediatric Respiratory Reviews 1 (3), 287–293.

Subcommittee on Energy and Environment, 2010. US House of Representatives Committee on Energy and Commerce, Hearing of 2/25/2010. <http://energycommerce.house.gov/index.php?option=com_content&view=article&id=1908:endocrine-disrupting-chemicals-in-drinking-water-risks-to-human-health-and-the-environment&catid=130:subcommittee-on-energy-and-the-environment&Itemid=71>.

Subcommittee on Health, 2010. US House of Representatives Committee on Energy and Commerce, Hearing of 4/22/2010; Hearing Transcript at pages 79 & 80. <http://energycommerce.house.gov/index.php?option=com_content&view=article&id=1964:the-environment-and-human-health-the-role-of-hhs&catid=132:subcommittee-on-health&Itemid=72>.

US EPA, 1998. Endocrine Disruptor Screening Program; Proposed Statement of Policy. Federal Register, vol. 63, No. 248, pp. 71541–71568, December 28, 1998. <http://www.epa.gov/scipoly/oscpendo/pubs/122898frnotice.pdf> (accessed 18.07.10).

US EPA, 2009. Environmental Protection Agency Endocrine Disruptor Screening Program (EDSP); Announcing the Availability of the Tier 1 Screening Battery and Related Test Guidelines. Federal Register, vol. 74, No. 202, pp. 54415–54422, October 21, 2009. <http://www.regulations.gov/search/Regs/contentStreamer?objectId=0900006480a4732danddisposition=attachmentandcontentType=html> (accessed 18.07.10).

US EPA, 2010. Weight-of-evidence Guidance Document: Evaluating Results of EDSP Tier 1 Screening to Identify Candidate Chemicals for Tier 2 Testing. Draft for Public Comment. <http://www.regulations.gov/search/Regs/contentStreamer?objectId=0900006480b80c60&disposition=attachment&contentType=pdf>. See also Federal Register (vol. 75, No. 213, pp. 67963–67965).

Weed, D.L., 2005. Weight of evidence: a review of concept and methods. Risk Analysis 25, 1545–1557.

Zaza, S., Wright-De Agüero, L.K., Briss, P.A., Truman, B.I., Hopkins, D.P., Hennessy, M.H., et al., 2000. Data collection instrument and procedure for systematic reviews in the guide to community preventive services. Task force on community preventive services. American Journal of Preventive Medicine 18 (Suppl. 1), 44–74.

Overview of Weight of Evidence Frameworks (Section II expanded)

While the various Weight of Evidence (WoE) frameworks that have been published were developed to fulfill different purposes, they discuss important principles and considerations for weighing evidence relevant to evaluating EDSP data. A recent critical evaluation of WoE methods describes a number of deficiencies that are important to avoid when developing a WoE framework (Krimsky, 2005). Krimsky observed that while published WoE methodologies claim to enhance clarity and transparency of evaluations, increase the consistency of regulatory decision-making, and identify the underlying assumptions, most WoE evaluations fail to accomplish those goals. The primary reason is that the epistemic foundation of WoE approaches is typically left undefined, which reduces clarity and consistency, and arguably undermines scientific integrity and validity. Krimsky further explains that a priori assumptions about the value of different evidentiary modalities are typically based on expert judgments rather than empirical facts, a factor that would seem to reduce scientific objectivity. Even the basis for making expert judgments about the relative quality and value of various types of evidence is typically not explained. Such expert judgments are then extended to produce "yes/no/ or maybe" determinations that require transforming continuous data, rich in complexity and biological detail, into simple dichotomous or triadic variables devoid of both context and clarity about the method of their derivation. In short, Krimsky finds that WoE conclusions often emerge from a black box of scientific judgments short on transparency and long on subjectivity.

The need for transparency in WoE determinations was underscored in a consensus opinion of a multi-sector committee recently convened by the International Life Sciences Institute's (ILSI) Health and Environmental Sciences Institute (HESI) (Schreider et al., 2010). They describe that when WoE determinations are criticized for lacking a clear scientific basis, the criticism is often due to a lack of clarity regarding what information was considered, the limited availability of the underlying scientific data, and too few details on the methodology used to arrive at a particular decision. They posit that the credibility of the data, the credibility of a risk assessment WoE conclusion, and the credibility of decisions based on the risk assessment are all highly dependent upon the transparency of the process at every level, and explain that credibility cannot be recovered at latter levels of the process once transparency has been breached. Schreider et al. (2010) and others (e.g., Borgert, 2007a,b) observe that the availability of raw data would go far toward enhancing the transparency and hence the credibility of published science.

The approach should also specifically account for certain characteristics of the system to be evaluated and the program devised to evaluate it. These characteristics include the homeostatic nature of the endocrine system, which is replete with multiple adaptive and compensatory signals that portend no potential for adverse effects. Distinguishing normal homeostatic, adaptive, and compensatory responses from those that may lead to adverse effects is implicit in Tier 2 of the U. S. EPA's EDSP, but cannot be addressed in Tier 1 due to limitations of the ESB. A WoE approach should also address the intentional redundancy of the ESB assays and endpoints, the intentional bias of the ESB toward avoiding false negative results at the expense of increasing the incidence of false positives, and the varying degree to which the ESB assays have been validated and their positive and negative predictive values ascertained. These and other issues related to the Tier 1 ESB are discussed in detail elsewhere (Borgert et al., 2011).

The European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC) framework for identifying endocrine active chemicals, in particular, offers insightful and useful decision matrices and considers the incorporation of data from the U.S. EPA's Tier 1 ESB (Bars et al., 2011; ECETOC, 2009). It emphasizes that every potential interaction of a chemical with endocrine pathways will not necessarily become manifest as endocrine activity in living organisms or lead to adverse effects. Because of this, chemicals that exhibit endocrine potential in the Tier 1 ESB would not all be expected to pose a hazard under actual conditions of use and exposure. Therefore, an element or assessment of potency is critical to a valid WoE assessment (Bars et al., 2011). Our approach incorporates a specific means of addressing this important consideration with a response weighting based on empirical results of various assays, described further in Section IV of the published manuscript.

ECETOC also explains that identifying a compound with potential endocrine active properties requires the analysis of regulatory (eco)toxicity data coupled with an understanding of the mode of action underlying the toxicity findings (Bars et al., 2011; ECETOC, 2009). The differences between apical endpoints and mechanistic endpoints indicative of particular modes of action, both of which are measured in various endocrine screening assays and reproduction and developmental toxicity tests, determines the interpretations that may be based on these different types of endpoints (Bars et al., 2011; Borgert et al., 2011; ECETOC, 2009). Therefore, our approach allows for differential quantitative weighting of these various endpoints depending on the nature of the endpoint and its specificity for evaluating any particular hypothesis. More mechanistic endpoints may be appropriately ascribed greater weight for hypotheses related to particular hormonal pathways, whereas apical endpoints may be ascribed more weight for hypotheses related to adverse outcomes. These considerations, however, are nuanced and must be clearly articulated as premises in the WoE evaluation.

In addition, we have incorporated specific principles described by the ECVAM workshop on WoE validation (Balls et al., 2006). These relate to the clear exposition of methods for searching and collecting relevant information to be used in the WoE determinations. We specifically incorporate the recommendations for GLP quality control and data availability, noting that there should be no preference for peer-reviewed journal sources per se, but merely for the type of accessibility they offer.

Building the Weight of Evidence Framework on Scientific Principles (Section III expanded)

Hypothesis Testing

We contend that each WoE determination requires a clearly-defined, testable hypothesis. Clear hypotheses afford the analyst an opportunity to estimate the potential error or uncertainty inherent in the conclusion, derived from a consideration of the experimental error and the degree to which experimental conditions and extraneous influences on the measurements can be controlled. From these latter parameters arises the possibility for alternative explanations of the data, which increases transparency of the inferences and conclusions. Without clear hypotheses, the logical connections between experimental data and deductive reasoning are undefined. The result is that any particular conclusion about a chemical is ambiguous and the transparency of inferences based on the data is greatly reduced. Thus, we advocate adopting a hypothesis-driven approach to WoE evaluations for endocrine activity, as

has been recommended and used previously (Goodman et al., 2006, 2009; Rhomberg 1998; Rhomberg & Goodman, 2008).

In order to be maximally testable and transparent, each WoE determination [a] through [e] listed in Section I of the published manuscript may best be subdivided into a series of specific hypotheses that can be tested with specific experiments or assays. This promotes use of a discrete set of endpoints for assessing each hypothesis, which should enhance the consistency of evaluations and interpretations. It further allows one to define the limits of interpretation of each assay against known positive and negative control chemicals for each hypothesis.

Primary Validity: Minimal Epistemic Status

To be valid and reliable, WoE determinations must consider the overall quality of data used in the evaluation. First, the minimal epistemic status, or 'primary validity' of the data should be considered. This involves evaluating results of each study according to minimum tenets of scientific validity, which have been articulated in a series of commentaries and editorials by Dr. Gio B. Gori, formerly Deputy Director of the Division of Cancer Cause and Prevention at the National Cancer Institute (Gori 1999, 2001, 2002, 2009a, 2009b, 2010). Gori explains that for data to be considered established scientific facts, they must, at a minimum, conform to three tenets underpinning the basic language of science that enables trustworthy measurement of the natural world. First, the identity and authenticity of scientific measurements must be verifiable within a defined range of precision. Second, measurements and observations must not be confounded by extraneous factors and influences known to corrupt their accuracy and precision. Third, the measurements and observations must be replicable in independent hands. These three tenets are undeniable and agreed to be the minimum requirements for valid regulatory science in the U.S. (Subcommittee on Energy and Environment, 2010; Subcommittee on Health, 2010). We believe they are also sufficiently unambiguous to provide the primary standard against which all data should be judged for use in WoE evaluations.

Although disarmingly simple, these three tenets are critically important and powerfully discriminative. To illustrate their use, consider vitellogenin production in male fish as an example, as it is one of the most frequently cited examples of environmental endocrine disruption. One would first determine whether a study measures what it purports to measure within a defined range of precision. The most obvious benefit provided by this first tenet is that it forces discrimination between the measurement itself and the interpretation often placed on the measurement. Here, the measured parameter is vitellogenin, usually in blood plasma but possibly in other organs such as liver. Vitellogenin, a dimeric glycolipophosphoprotein, is the egg yolk precursor protein in all oviparous vertebrates and can be measured by a variety of techniques (e.g., Alda & Barceló, 2001; Wheeler et al., 2005; Wu et al., 2006), each of which is verifiable within defined margins of error. However, causal links between this protein and reproductive impairment or population level effects have not been established, and in the absence of causal links, measurement of vitellogenin in male fish cannot be equated with "endocrine disruption" (Mills et al., 2003; Mills & Chichester, 2005). Establishing causality would require further experimental evidence utilizing counterfactual study designs. Second, the experimental conditions under which the measurements are taken may be difficult to control. Besides methodological controls related to the particular analytical technique, measurements of plasma vitellogenin in male fish would need to consider background levels of the protein within the study population as well as the presence of viruses that have been shown to affect plasma vitellogenin levels in both male and female fish (Trubiroha et al., 2010), and perhaps other factors depending upon whether the study is conducted in the laboratory or the field. Finally,

one would consider whether the measurements have been repeated in independent hands, meaning in different laboratories not affiliated with one another or by different investigators not employing the same instruments and personnel. As evidenced by the publications cited here, measurement of plasma vitellogenin in male fish is generally repeatable in independent hands given adequate experimental and methodological controls and similar study designs.

Secondary Validity: Data Reliability and Transparency

WoE evaluations should consider the reliability of the reported data, which may also be called 'secondary validity'. Klimisch et al. (1997) have defined reliability in terms of the transparency and thoroughness of data reporting. For in vivo studies, they advocate giving greater weight to studies that report detailed information on the test species, test substances (purity, origin), number of animals evaluated, scope of investigations per animal (e.g., clinical chemistry, organ weights, hematology, histopathology), description of changes or lesions observed, control and historical control groups, test conditions, route of administration, dose schedule and dose concentration (including analytical verification). For in vitro studies, Klimisch et al. assign greater weight to studies that report similar details regarding the test substance, but also information specific to in vitro assays such as the test system and method, positive and negative controls, interferences with the method, data on secondary effects that can influence the result (solubility, impurities, pH shifts, osmolarity, etc.), and they request similar information of ecotoxicity studies, as well as information on the life stages of the animals studied. Based on evidence of potential confounding in studies of endocrine activity, specific additions to Klimisch's list could be considered, e.g., composition of diets, composition of water bottles and cage materials, bedding, stressors such as handling and manipulation, and any other factors that could affect hormonal systems, as well as details on the mathematical and statistical algorithms used to analyze the data.

Klimisch et al. (1997) note that studies conducted under Good Laboratory Practices requirements (GLP) in accordance with regulatory guidelines ("guideline studies") are required to record all such information and recommend that such studies be used as reference standards for the evaluation of reliability. We agree with this premise, but add the proviso that to serve as reference standards, guideline studies should have either been subjected to thorough validation programs such as those required under ICCVAM or ECVAM[1] or have been used extensively so that their performance is well characterized. Klimisch et al. do not constrain their highest reliability rating (code) to the results of guideline studies, but concede that any study reporting adequately on these parameters should be given greater weight than studies lacking them, irrespective of whether the study is conducted according to a regulatory guideline or GLP. We concur that studies with greater thoroughness, transparency and availability of recorded data should be considered more reliable, irrespective of the source or venue under which a study was conducted. Application of uniform, objective criteria, such as described by Schneider et al., (2009) provides a scientifically sound basis for assigning appropriate weights to all relevant toxicity studies, both GLP and non-GLP.

Although the U.S. EPA's endocrine screening battery comprises guideline studies, and while some of these guidelines, such as the uterotrophic assay, have been subjected to extensive validation programs, others fall well short of meeting the requisite degree of sensitivity and

1 Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) and European Centre for Validation of Alternative Methods (ECVAM)

specificity required for confidence that such methods can truly differentiate a non-endocrine active agent from an active one. Furthermore, although their use within the pharmaceutical industry verifies that the binding assays and the uterotrophic and Hershberger assays have proven ability to identify chemicals with potent hormonal activity, it remains to be verified that these ESB assays are able to distinguish chemicals that are weakly hormonal in vivo from false positives, i.e., chemicals that produce a signal in the assay but never manifest the implied hormonal activity in vivo. For this reason, the Tier 1 ESB cannot be ascribed the same level of reliability as other guideline studies until more extensive validation data are available (Borgert et al., 2011). The U.S. EPA has issued test orders for conducting the screening battery on 67 pesticide chemicals, many for which extensive reproductive and developmental toxicity data are already available. Depending on the clarity of the results, the initial round of screening might be viewed as a step toward validation because the predictive value of the screens can then be estimated by comparison to the results of definitive guideline studies for these chemicals.

Tertiary Validity: Relevance and Probative Power of Study Design and Causality

Regardless of the plausibility that an adverse effect is the result of endocrine activity, demonstrating that a specific endocrine mode of action underlies a particular adverse effect requires more than the mere coincidence of a Tier 1 screening result with an effect in Tier 2 testing. Hence, the probative power of the study design must be evaluated, a feature that may be deemed 'tertiary validity'. Establishing a causal relationship between endocrine activity and adverse outcome is essential. Although this principle is germane to other determinations as well, it applies most directly to WoE determinations [c] – determining from the results of Tier 2 toxicity tests whether a substance exhibits adverse effects potentially mediated by androgen, estrogen, or thyroid pathways, and [d] – determining from Tier 1 ESB, OSRI, Tier 2 toxicity tests, and as necessary, additional mode-of-action experiments, whether the adverse effects observed in Tier 2 toxicity tests are a consequence of endocrine activity. In this regard, an initial establishment of biological plausibility is required. Because a great deal is known about the steps involved in eliciting effects in each of the Tier 1 ESB assays, positive results can be used to develop an initial hypothesis of the potential mechanism of action. Because all steps will not be known with certainty, it is appropriate to describe this as a "working hypothesis of the mode of action." This will allow one to draw reasonable working postulates concerning the agent's influence on key processes. Following the development of working hypotheses, causal relationships between the proposed mechanistic steps should be verified using counterfactual experimental methods. Counterfactual concepts and approaches for establishing causality in pharmacology, toxicology, and epidemiology have recently been critically evaluated (Borgert & Gori, in prep) and these concepts are explicitly incorporated here.

References Cited in Supplemental Materials not cited in main manuscript

Alda, M. J., Barceló, D., 2001. Review of analytical methods for the determination of estrogens and progestogens in waste waters. Fresenius' Journal of Analytical Chemistry. 371 (4), 437-447.

Mills, L. J., Gutjahr-Gobell, R. E., Horowitz, D. B., Denslow, N. D., Chow, M. C., Zaroogian, G.E., 2003. Relationship between reproductive success and male plasma vitellogenin concentrations in cunner, *Tautogolabrus adspersus*. Environmental Health Perspectives. 111 (1), 93-100.

Mills, L. J., Chichester, C., 2005. Review of evidence: Are endocrine-disrupting chemicals in the aquatic environment impacting fish populations? The Science of the Total Environment. 343 (1-3), 1-34.

Trubiroha, A., Kroupova, H., Wuertz, S., Frank, S. N., Sures, B., Kloas, W., 2010. Naturally-Induced endocrine disruption by the parasite *Ligula intestinalis* (Cestoda) in roach (*Rutilus rutilus*). Gen Comp Endocrinol. 166 (2), 234-40.

Wheeler, J. R., Gimeno, S., Crane, M., Lopez-Juez, E., Morritt, D., 2005. Vitellogenin: A review of analytical methods to detect (anti) estrogenic activity in fish. Toxicol Mech Methods. 15 (4), 293-306.

Wu, C., Yuan, D., Liu, B., 2006. Rapid determination of vitellogenin in fish plasma by anion exchange high performance liquid chromatography using postcolumn fluorescence derivatization with o-phthalaldehyde. Anal Sci. 22 (12), 1593-6.

Table A. Assays and Endpoints for Use in Hypotheses [a]-1 through [a]-8[1]

Tier 1 Assay	Endpoints	Prototypical Response of Natural Hormone or Surrogate for Natural Hormone	Response[2] of Negative Control	ESB Response[2] of Test Agent W_{RES}	OSRI Response[2] of Test Agent W_{RES}	Relevance Weighting[3] for Endpoint W_{REL}
ER RBA	ER agonism/antagonism calculated as an IC50 value					
ER TAA	ER agonism calculated as PC10, PC50 and EC50 as appropriate					
AR RBA	AR agonism/antagonism calculated as an IC50 value					
Aromatase	Aromatase inhibition calculated as an average IC50 value					
Steroidogenesis	Estradiol Levels Testosterone Levels					
Uterotrophic	Statistically significant increased uterine weight at high dose and uterine weight > 130% of control Dose – dependent increase in uterine weight					
Hershberger	Cowper's Gland Weight Testicular Weights Seminal Vesicle Weight Levator Ani-bulbocavernosus Muscle Weight Glans Penis Weight					

Intact Male [4]	Testes Weight Epididymides Weight Prostate Weight Seminal Vesicle Weight Accessory Sex Gland Weight (prostate plus seminal vesicles with coagulating gland) Thyroid Weight Testes histopathology Thyroid histopathology Epididymides histopathology Hormone Concentration Response profile					
Pubertal Male	Growth (daily body weight) Age and Weight at Preputial Separation Seminal Vesicle + Coagulating Gland Weight Ventral Prostate Weight Dorsolateral Prostate Weight Levator Ani + Bulbocavernosus Muscle Complex Weight Epididymis Weight Testes Weight Thryoid Weight Adrenal Weight Pituitary Weight Blood Chemistry, standard panel Hormone Levels Testes histopathology Thyroid histopathology Epididymides histopathology					
Pubertal Female	Growth (daily body weight) Age and Weight at Vaginal Opening Uterus (blotted) Weight Ovaries (paired) Weight Thyroid Weight Pituitary Weight Adrenals (paired) Weight Uterus histopathology Ovary histopathology Thyroid (colloid area and follicular cell height) histopathology Blood Chemistry, standard panel Hormone Levels Estrous Cyclicity					
Fish Screening Assay	Vitellogenin Secondary Sexual Characteristics Fecundity Sex Steroid Concentration Gonad Histopathology Gonad Somatic Index Behavior Fertilization Success Adult Survival					

AMA	Advanced development Asynchronous development Thyroid histopathology Delayed development Snout-Vent Length Hind Limb Length Wet Weight Survival Sublethal Observations					
Endpoints from OSRI (Other Scientifically Relevant Information)						

[1.] This table would be repeated for each hypothesis [a]-1 through [a]-8.
[2.] To include consideration of dose response, statistical significance and biological significance.
[3.] Obtain W_{REL} from consensus workshop. W_{REL} values for the various endpoints would be expected to differ depending upon the hypothesis under evaluation. Some endpoints might have no relevance for a particular hypothesis.
[4.] The intact male assay is not included in the U.S. EPA's Tier 1 ESB, however, it has advantages that may warrant conducting it to improve the overall interpretability of Tier 1 (Borgert et al., 2011).

Table B. Summary of Weighting Scores for Hypotheses [a]-1 through [a]-8[1]

Tier 1 Assay	Endpoints	Composite Relevance Weighting for Assay	Composite Response Weighting from All Endpoints	Overall Weighting for Assay $\sum(W_{REL} \times W_{RES})^2$
ER RBA				
ER TAA				
AR RBA				
Aromatase				
Steroidogenesis				
Uterotrophic				
Hershberger				
Intact Male[3]				
Pubertal Male				
Pubertal Female				
Fish Screening Assay				
AMA				
Other Scientifically Relevant Information(OSRI)				

[1] This table would be repeated for each hypothesis [a]-1 through [a]-8.
[2] The sum of the products for each endpoint measured in the assay.
[3] The intact male assay is not included in the U.S. EPA's Tier 1 ESB, however, it has advantages that may warrant conducting it to improve the overall interpretability of Tier 1 (Borgert et al., 2011).