Title: Endocrine Disruption: Where have we been, interpretation of data, and lessons learned from Tier 1

Running head: Lessons learned from Tier 1

Jane P. Staveley *[†], Leslie W. Touart[‡], Keith Solomon[§], Ellen Mihaich[|], Amy Blankinship[#], and Gerald Ankley[††]

†Exponent, 1000 Centre Green Way, Suite 200, Cary, NC, USA, 27513; Telephone: 919-228-6480; Fax: 919-228-6501; e-mail: jstaveley@exponent.com; [‡] USEPA, U.S. EPA, Office of Chemical Safety and Pollution Prevention, Washington, DC, USA, [currently Equiparent Consulting, Woodbridge, VA, USA, les.touart@equiparentconsulting.com]; [§]Centre for Toxicology, University of Guelph, Guelph, Ontario, Canada, ksolomon@uoguelph.ca; [|]Environmental and Regulatory Resources, Durham, NC, USA, emihaich@nc.rr.com; [#]U.S. EPA, Office of Pesticide Programs, Washington, DC, USA, blankinship.amy@epa.gov; [††]U.S. EPA, Office of Research and Development, Duluth, MN, USA, ankley.gerald@epa.gov

* To whom correspondence may be addressed

Corresponding Author:

Jane Staveley

1000 Centre Green Way, Suite 200, Cary, NC, USA 27513.

Email address: jstaveley@exponent.com

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

ABSTRACT

In response to the requirements of the US EPA's Endocrine Disruptor Screening Program, Tier 1 assays have been performed with a number of pesticides over the past several years. These assays are designed to be used in concert as a screen for potential interactions with vertebrate estrogen, androgen, and thyroid systems. The results of the 11 assays in the Tier 1 battery are then used, along with other lines of evidence, to determine whether a chemical is endocrine-active and, as a consequence, might be a candidate for Tier 2 testing. An overview of the Tier-1 testing program was presented in Session Two of the Society of Environmental Toxicology and Chemistry (SETAC) North America Focused Topic Meeting: *Endocrine Disruption Chemical Testing: Risk Assessment Approaches and Implications* (February 4 – 6, 2014). Subsequent presentations discussed the concept of weight-of-evidence (WoE) and assessment of Tier 1 results in a WoE framework. The importance of scientifically credible, transparent approaches for conducting WoE analyses was recognized, and approaches for framing the hypotheses, evaluating the data, assigning weight to different endpoints relative to their diagnostic effectiveness, and assessing confounding factors were presented. In recognition of the cross-species conservation of the hypothalamic-pituitary-gonadal axis among vertebrates, a subset of the Tier-1 *in vivo* assays may be useful for more rapidly screening chemicals for potential endocrine activity.

Keywords: Endocrine disruption, Endocrine Disruption Screening Program, testing, weight-of-evidence

INTRODUCTION

Session Two of the Society of Environmental Toxicology and Chemistry (SETAC) North America Focused Topic Meeting: *Endocrine Disruption Chemical Testing: Risk Assessment Approaches and Implications* (February 4 – 6, 2014) focused on the experience gained to date with implementation of the Tier 1 testing of U.S. EPA's Endocrine Disruptor Screening Program (EDSP), and how these data can be used to make decisions about the need for further testing. Leslie Touart presented an overview of the 11 assays in the Tier 1 screening battery. Keith Solomon discussed the concept of using weight-of-evidence (WoE) in risk assessment, illustrated by an example on the potential effects of atrazine on fish, amphibians, and reptiles. Ellen Mihaich described a hypothesis-based weight of evidence framework that was developed to evaluate experimental data, with a proposed specific use in evaluating results of the Tier 1 screening battery. Amy Blankinship provided an overview of the conceptual basis of the WoE guidance used by the USEPA to evaluate Tier 1 data for identifying the need for additional (Tier-2) testing. The session concluded with a presentation by Gary Ankley on an analysis indicating that it appears possible to use just two of the current Tier-1 tests as initial "gate keeper" assays, following which chemicals may be exempted from further testing or subjected to additional, confirmatory analyses with other existing Tier-1 assays.

SESSION PRESENTATION SUMMARIES

USEPA Endocrine Disruptor Screening Program (EDSP) Tier-1 Battery Overview by: Leslie Touart

The suite of 11 Tier-1 EDSP assays is specifically designed to detect chemicals with the potential to interact with the estrogen, androgen, and thyroid (EAT) systems in vertebrates,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

through mechanisms such as activation and antagonism of target nuclear hormone receptors, and inhibition of hormone synthesis (http://www.epa.gov/endo/). Given the complex interactive nature of the endocrine system, if the objective is to comprehensively detect their potential to disrupt endocrine regulated processes, it is clear that chemicals should be tested for apical effects (*e.g.*, the ability to alter growth, development, or reproductive processes) and their potency in *in vitro* assays of receptors and synthesis of sex steroids. A battery of screening tests has been developed which includes a range of taxonomic groups and sufficient diversity of endpoints to maximize sensitivity and minimize false negatives. There are five *in vitro* assays focused on binding to and transactivation of the estrogen receptor, binding to the androgen receptor, and inhibition of synthesis of sex steroids. There are six *in vivo* Tier-1 screens, four utilizing rats (uterotrophic and Hershberger assays; male and female pubertal assays), one with the fathead minnow (fish short-term reproduction assay; FSTRA), and one with the amphibian *Xenopus laevis* (amphibian metamorphosis assay; AMA). Although each of the Tier-1 assays provides unique data, the suite was purposefully designed to result in some redundancy with respect to detecting endocrine pathways of concern (Table 1). The *in vitro* assays provide sensitivity and mechanistic clues, while the *in vivo* assays provide for integrative responses and metabolism and distribution considerations. The results of the Tier-1 battery are to be interpreted in a WoE context, rather than the sum of positive and negative assays. Some endpoints are more diagnostic/specific than others, and effects seen in multiple endpoints and multiple assays carry the most weight. There are two possible interpretations of the outcome of the Tier-1 battery: either the potential for EAT activity exists, which warrants analysis in Tier-2 testing, or there is low or no potential for EAT activity. A FIFRA Science Advisory Panel meeting held in 2008 to review the Tier-1 screening battery concluded that, based on the state of the science at the time,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

the set of assays were an appropriate starting point to detect potential endocrine disruptors and should continue to be refined and developed. In summary, multiple assays are required to comprehensively screen endocrine, androgen, and thyroid hormone systems. The *in vitro* assays are suitable for well-understood mechanisms (*e.g.*, receptor binding), while the *in vivo* assays with intact hypothalamic-pituitary-gonadal (HPG) and hypothalamic-pituitarty-thyroidal (HPT) axes are useful for efficiently screening complex processes. The totality of the results of the Tier-1 screening battery are needed to support WoE conclusions about the potential of a chemical to interact with vertebrate EAT systems.

Use of weight of evidence for characterizing adverse outcome pathways in risk assessment by: Keith Solomon

Information and data on chemicals from studies published in the open literature are increasingly being used for assessment purposes by regulatory agencies in many jurisdictions, including North America and Europe. Because most of these studies are not conducted to the Good Laboratory Practices (GLP) standards as required by regulatory agencies, there is a need to assess their quality and relevance in light of the regulatory endpoints being considered. To aid in interpretation and to use these data in regulatory decision-making, they need to be integrated into lines of evidence that inform adverse outcome pathways (AOPs) and lines of evidence related to apical endpoints such as survival, growth, development, and reproduction.

There are important differences between studies published in the open literature and those conducted under GLP guidelines for regulatory agencies. Published studies often are incompletely documented, raw data are rarely available, and many studies do not follow standardized protocols. In addition, studies used in reviews and meta-analyses may be subjected

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

111 to selection bias or, in a worse case, there may be selection bias where negative (no effect)
112 results are not published (Walker et al. 2008). In contrast, studies conducted under GLP with
113 Quality Assurance and Quality Control (QA/QC) are required by regulation in many
114 jurisdictions, are completely documented, the raw data are available, most studies are conducted
115 using standardized protocols, and there is no publication bias; all observations are documented.
116 For this reason, GLP studies with QA/QC cannot always be directly compared to or combined
117 with published studies for a WoE analysis for decision making. A WoE analysis of the potential
118 effects of atrazine on fish, amphibians, and reptiles (Van Der Kraak et al. 2014) was conducted
119 using quantitative methods to characterize the strength and relevance of published and GLP
120 studies. This brief overview describes a subset of data taken from Van Der Kraak et al. (2014)
121 with a specific focus on reproductive outcomes in fish, amphibians, and reptiles.
122 In this example, the strength of the experimental methods and the ecological relevance of
123 the observed responses from over 2000 studies and experiments were scored. The detailed
124 methods of scoring are reported (Van Der Kraak et al. 2014) and are not repeated here. Briefly,
125 the strength of the methods was scored based on various aspects of the studies, such as the
126 experimental design and conduct, the use of appropriate controls, measures of exposures, the
127 inclusion of environmentally realistic concentrations, number of concentrations, quality control,
128 and transparency of data. These criteria are similar to those suggested by Klimisch et al. (1997).
129 The relevance of the each response was assessed by scoring statistical significance, concentration
130 or dose-response, its relevance to an appropriate apical endpoint, and a biologically plausible
131 mechanism. The WoE process was inclusive and no studies were excluded, except those with
132 mixtures where the individual components were not tested individually. Results were presented
133 graphically where strength and relevance were shown separately for easy interpretation and are

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

supported with details of the experimental procedures (see SI provided with Van Der Kraak et al. 2014).

AOPs (Ankley et al. 2010) are used to characterize links between responses at lower levels of biological organization and apical endpoints such as survival, growth, development, and reproduction (Figure 1). AOPs provide the framework for extrapolation of effects to other organisms/taxa or to identify reliable and robust biomarkers that can be used in place of the apical endpoint. Responses from multiple studies at each level of an AOP can be subjected to WoE analysis. If one or more apical endpoints (4 and 5 in Figure 1) have been characterized under WoE, and the combination of these indicates no or *de minimis* effects at environmentally relevant exposures, an analysis of AOPs is not needed. In this case any effects observed at lower levels of organization are "trumped" or negated by lack of effect on apical endpoints and those at lower levels are most likely only bioindicators of exposure or adaptive response. However, if one or more of the apical endpoints indicates relevant effects at environmentally-relevant exposures, then a characterization of AOP might be useful to better understand the response. Because responses in an AOP are concatenated, a break in the chain at any point in the pathway (illustrated by the red X in Figure 1) provides evidence that the responses are not important for apical effects and that regulatory action would not be needed.

To illustrate the combination of AOPs with WoE analysis, reproductive responses to atrazine in fish, amphibians, and reptiles were combined in graphs showing the mean scores for multiple responses and their uncertainty (see Van Der Kraak et al. (2014) for details) in four links of an AOP. These links in the AOP chain were at the biochemical (A), cellular (B), organ (C), and organism (D) levels (Figure 2). The organismal level is apical. As can be seen from the graphics in Figure 2 (A to D), the mean values for relevance of all the responses in the AOP

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

chain cluster at the low end of the relevance scale. The means and uncertainty of the scores provide the basis for testing risk hypotheses in the WoE framework. These analyses suggested that there was a *de minimis* risk of adverse effects at all levels of the AOP. Strictly speaking, the lack of effects at the organismal level would negate the need for AOP analysis but the example is illustrative of the robustness of the response as effects at all levels of the AOP are of low relevance. This provides greater assurance that the lack of response is real and not just due to a lack of data or measures at different levels of organization. As is indicated by the error-bars (Figure 2), there was less uncertainty in the scores for relevance than the scores for strength. The scores from strength for these responses (see details in Van Der Kraak et al. 2014) ranged from low to high but the high-strength scores were consistent in indicating very low or *de minimis* relevance.

In conclusion, the use of a formal, well described, transparent, and quantitative process for WoE provides a helpful tool for conducting risk assessment. It is more objective and, when combined with analysis of AOPs, provides more clarity and understanding of the significance of effects. The example provided is directed specifically to reproduction but the process is applicable to areas other than risk assessment; however, different and response-specific methods of scoring may be needed.

"Weighing" the Evidence: Relevance and Transparency in the Evaluation of Endocrine Activity by: Ellen Mihaich

A comprehensive, hypothesis-based weight of evidence (HB-WoE) framework was developed to be applicable to any determination relying on experimental data, with a proposed specific formulation for evaluating results of the U.S. EPA's Tier-1 Endocrine Screening Battery

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

(ESB) (Borgert et al. 2011a). The framework requires that before any WoE determinations are considered, each experimental endpoint be weighted according to its relevance for deciding each of 8 hypothesis addressed by the ESB. These hypotheses test whether or not the chemical under evaluation has the potential to act as an (anti)-estrogen, (anti)-androgen, (anti)-thyroid, or induce or inhibit steroidogenesis. The purpose of an *a priori* relevance weighting is to ensure a level of transparency and objectivity exceeding that possible from WoE processes claiming a basis in professional judgment alone. Ideally, quantitative relevance weighting (Wrel) values would be derived from data revealing the positive and negative predictive value of the various endpoints for the hypotheses addressed by the ESB assays. Because the ESB assays have not been validated to that level (Borgert et al. 2011b), obviating the derivation of quantitative Wrels, this method provides for endpoints to be ranked according to 4 categories based on interpretations of relevant literature (Borgert et al. 2014). Although these Wrel rankings necessarily involve professional judgment, their *a priori* derivation based on a defined rationale (Borgert et al. 2014) enhances transparency nonetheless and renders any WoE determinations based on them amenable to methodological scrutiny according to basic scientific premises. To make WoE determinations for a particular substance, the framework requires combining Wrel values/rankings for each hypothesis with response weightings (Wres) derived from the ESB data.

The method has been more fully described by Borgert et al. (2014). Wrels were determined by ranking the endpoints by hypothesis according to the following definitions below. Although no hypothesis can be decided on the results of a single assay, "interpretable" means that the results for an endpoint provide information relevant to the hypothesis, without clarification from other endpoints. Whether a hypothesis is supported requires consideration of results from all relevant (#1, #2, #3) assays and endpoints. Rank 1 endpoints are typically *in vivo*

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

endpoints, specific & sensitive for the hypothesis and interpretable without other endpoints. Rank 2 includes many *in vitro* endpoints that are sensitive and specific, but less informative than Rank 1. Rank 3 includes many apical *in vivo* endpoints that are relevant for the hypothesis, but are only corroborative of Rank #1 and #2 endpoints. Rank 4 endpoints were considered not relevant for the hypothesis.

Data for the test chemicals are evaluated for each hypothesis individually, beginning with Rank 1 and continuing through Rank 3 endpoints. The response to Rank 1 endpoints guides the evaluation and interpretation of information from lower-ranked endpoints. Responses in Rank 1 are a preliminary indication that the hypothesis is or is not supported. Rank 2 endpoints are then evaluated, with consistent positive responses among Rank 1 and 2 endpoints considered sufficient support, and consistent negative responses considered refutation of the hypothesis. Rank 3 endpoints are then consulted for consistency and, together with the strength of response (Wres) in Rank 1 and 2 endpoints, temper or strengthen the conclusion. The interpretation becomes more complex if Rank 2 endpoints are inconsistent with negative results in Rank 1 endpoints. In this case, the strength of the response in Rank 2 endpoints becomes even more critical, as does an evaluation of Rank 3 endpoints, along with a consideration of the potential reasons that Rank 1 endpoints might not respond. Some overarching guidelines for interpretation can be established. Rank 1 endpoints cannot be dismissed for inconsistency with Rank 2. Rank 3 endpoints, in contrast, provide little useful information other than as corroboration for findings in Ranks 1 and 2. Situations in which Rank 2 and 3 are consistent, but inconsistent with Rank 1 endpoints present the greatest challenge, and no general statements can be made.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

Published data from genistein was used to illustrate the application of this WoE framework and process for determining the potential for genistein to act as an estrogen agonist. Genistein is an isoflavone present in plant foods like soy, fava beans, and clover. Phytoestrogens like genistein are known to cause effects on reproduction in female ruminants, such as sheep and cattle (Adams 1995), and have been well studied to understand potential impacts on humans given the number of populations using a diet high in soy. For brevity, summary results are presented only for the estrogen agonist hypothesis in Table 2. In this example, although there are studies that provide some conflicting results (data not shown), the overall weight of the evidence of the data for genistein would support the estrogen agonist hypothesis. While few studies use positive controls because of animal use concerns, and specific positive controls would be needed to address each hypothesis being tested, some studies with genistein have employed compounds such as ethinyl estradiol (Kim et al. 2005) which allows for an estimation of estrogenic potency. Each additional hypothesis and the appropriately ranked endpoints would be considered separately; more detail on endpoint ranking can be found in Borgert et al. (2014).

This HB-WoE framework has been criticized for excessive detail, burdensome number and impossible requirements for quantitative rankings, and excessive time required to complete the process. As shown here, these criticisms are unfounded. The HB-WoE framework (Borgert et al. 2011a) provides a means for transparent, objective conclusions about ESB results, and moreover, streamlines the evaluation by allowing the analyst to appropriately allocate time and attention to the most definitive information. Although it is not yet possible to attain the goal of data-derived quantitative Wrel and Wres values, use of explicit Wrel rankings, derived *a priori* and applied similarly for each hypothesis, helps to ensure transparency and consistency, a feature absent from WoE approaches based solely on professional judgment. Despite an absence of

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

positive and negative control data in some ESB assays, Wres information can often be gleaned from Rank 1 and some Rank 2 endpoints, including an estimate of potency differences. The HB-WoE framework provides for efficient processing and interpretation of ESB data by considering the results of Rank 1 through 3 endpoints in consecutive order for each hypothesis. It provides for a systematic method for identifying and resolving inconsistencies in results from ESB and other scientifically relevant information and obviates a need to consider less definitive information unless it could help to resolve an ambiguous interpretation.

Weight of Evidence: Evaluating Results from Tier-1 Screening for the U.S. EPA Endocrine Disruptor Screening Program by: Amy Blankinship

In 2011, the United States Environmental Protection Agency's Office of Chemical Safety and Pollution Prevention (EPA/OCSPP) published a guidance document for the Endocrine Disruptor Screening Program (EDSP) which presented a weight of evidence (WoE) approach for evaluating Tier-1 screening data for identifying the need for additional (Tier-2) testing (USEPA 2011). The function of the EDSP Tier-1 screening process is to identify chemicals that have the potential to interact with the estrogen, androgen, or thyroid (E, A, or T) pathways and evaluate the need for additional testing. The WoE guidance document provides general guidance in support of EPA efforts to integrate and interpret data submitted in response to orders for Tier-1 screening; however, the guidance is not considered binding and reviewers may deviate from the guidance where circumstances warrant. As described in the guidance document, the WoE process identifies how the individual lines of evidence are assembled and integrated along two concepts (*i.e.*, complementarity and redundancy) within the conceptual framework of an adverse

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

outcome pathway (AOP). Broadly, there are four main steps outlined in the guidance which provide the foundation for WoE evaluations. The first step is to assemble and evaluate the individual studies for their scientific quality and relevance in evaluating potential endocrine interaction(s). The second step is to integrate the data along different levels of biological organization while examining the extent of concordance (robustness) of complementarity (*i.e.*, the concordance of endpoints within an assay that measures multiple endpoints) and redundancy (*i.e.*, the concordance of endpoints/responses across assays) in the observed responses across these different levels of biological organization. The third step is to then characterize the main lines of evidence as well as any conclusions. Finally, the last step is to evaluate whether additional testing is needed based on the evidence and conclusions described above.

As mentioned, the first step is to assemble and evaluate the available scientific data. Data for the EDSP Tier-1 WoE evaluation falls into one of two categories: 1) EDSP Tier-1 data, and 2) other scientifically relevant information (OSRI). The EDSP Tier-1 data represent a battery of 11 assays consisting of *in vitro* and mammalian and wildlife *in vivo* assays. OSRI may include published literature studies as well as studies conducted under USEPA (often referred to as Part-158 data) or OECD guidelines submitted in support of registration of pesticides or other chemicals. Each study is evaluated for scientific quality and relevance for informing interactions with the E, A, or T pathway. Additionally, the concordance or consistency (complementarity) of the responses in the individual study is evaluated. For the Tier-1 *in vivo* assays, often multiple endpoints are measured in each assay. Decision logic trees were developed for some Tier-1 *in vivo* assays in an effort to help guide the investigator/reviewer in interpreting results across multiple endpoints within an assay (Ankley and Jensen 2014; USEPA 2009). Evaluation of the potential confounding effects of overt toxicity in the study as well as the relative degree of

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

292 diagnostic utility of a specific endpoint for discerning whether or not the chemical has interacted
293 with the endocrine system are considered. The collective response of the individual endpoints, as
294 well as the conditions under which they were expressed, are considered when evaluating an
295 overall indication of potential interaction as measured by the study.

296 The second step in this WoE process is to formulate hypotheses and integrate the
297 available data along different levels of biological organization. Two key elements in the
298 integration of data as well as characterizing the extent to which the available data support a
299 hypothesis that a chemical has the potential to interact with E, A, or T pathways are the concepts
300 of complementarity and redundancy. These two concepts provide a basis for considering the
301 plausibility, coherence, strength, and consistency of the body of evidence. The current EDSP
302 Tier-1 screening assays are meant to evaluate whether or not a chemical can interact with E, A
303 and T consisting of different levels of biological organization from a molecular initiating event
304 such as receptor binding through potential adverse effects in apical endpoints such as sexual
305 development and fecundity at the whole organism level. Transitions to higher levels of biological
306 organization can indirectly provide information on potential compensatory capabilities of an
307 individual organism.

308 After the data have been assembled and integrated, the third step is to characterize the
309 main lines of evidence along with the conclusions; this characterization involves three
310 components. The first component is whether the data provide relevant, robust and consistent
311 evidence in terms of complementarity and redundancy as well as biological plausibility. Second,
312 is at what level of biological organization were the responses observed and whether organisms
313 exhibit compensatory responses at higher-levels of biological organization. Finally, an

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

evaluation of under what conditions did the responses occur including discussions regarding whether the responses were observed in the presence of overt or systemic toxicity. The presence of overt and/or systemic toxicity introduces uncertainty in the ability to distinguish effects specifically related to an endocrine-mediated effect from a non-endocrine toxic response. This uncertainty in distinguishing whether the responses were endocrine-mediated was discussed at the FIFRA Scientific Advisory Panel (SAP) meeting in July 2013 that evaluated scientific issues associated with the WoE evaluation of the EDSP Tier-1 screening process. The SAP stated that , "*In summary, the Panel agreed that little, if any, weight should be placed on signs of endocrine disruption in the presence of overt toxicity. All effects in endocrine sensitive tissues should be evaluated in terms of primary interactions with the endocrine system vs. secondary effects related to toxicity in non-endocrine organs or overall disruptions in homeostasis"* (Schlenk and Jenkins, 2013; Page 12; SAP 10/30/2013). Therefore, EPA considers multiple lines of evidence in including the observed responses in the Tier-1 assays and OSRI in the context of a chemical's physical/chemical properties and its known modes of action in its overall characterization of a chemical's potential to interact with the E, A or T pathway. Adequately addressing these three main questions is fundamental to the WoE process and in determining whether additional data are needed.

In addition to characterizing the WoE, reviewers also consider: 1) uncertainties and their potential impact to conclusions; 2) discussion of key studies; 3) description of inconsistent or conflicting data; 4) overall strength of evidence supporting a conclusion; and, 5) what, if any, additional data are needed and why. Assessing the need for additional data is based on a case-by-case analysis which will include integration of existing knowledge on the chemical including relevant hazard and exposure information. In summary, the evaluation of the EDSP Tier-1

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

screening process and ultimate decision for any additional testing is based on a totality of the scientific evidence.

Cross-Species Conservation of Endocrine Pathways Provides a Basis for Reevaluation of EDSP Tiered Testing Paradigm: by Gerald Ankley

Many structural and functional aspects of the HPG axis are known to be highly conserved, but the relative significance of this from a regulatory toxicology perspective has received comparatively little attention. High-quality data generated through development and validation of Tier-1 tests for the USEPA Endocrine Disruptor Screening Program (EDSP) offer a unique opportunity to compare responses of mammals versus fish to chemicals that may affect shared pathways within the HPG axis. The analysis described by Ankley and Gray (2013) focused on data generated with model chemicals that act (primarily) as estrogen receptor agonists (17α-ethynylestradiol, methoxychlor, bisphenol A), androgen receptor agonists (methyltestosterone, 17β-trenbolone), androgen receptor antagonists (flutamide, vincolozolin, p,p'-DDE) or inhibitors of different steroidogenic enzymes (ketoconazole, fadrozole, fenarimol, prochloraz). All 12 chemicals had been tested in the EDSP fish short-term reproduction assay (FSTRA) and in one or more of the four *in vivo* Tier-1 screens with rats (Uterotrophic, Hershberger, male and female pubertal assays). In most cases there was high concordance between the fish and rat assays with respect to identifying chemicals that impacted specific HPG pathways of concern, with the test chemicals producing positive results in the fish and one or more of the rat assays. However, some assays were clearly superior to others in terms of detecting specific pathways; for example, the effects of inhibitors of steroid hormone synthesis were most obvious in the FSTRA, whereas the activity of androgen receptor antagonists were

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

clearest in the Hershberger and male pubertal assays. Based on this analysis it appears possible to use just two of the current Tier-1 tests, the FSTRA and the male pubertal assay, to ensure full coverage of HPG axis pathways of concern. Specifically, these two tests could serve as initial "gate keeper" assays, following which chemicals may be exempted from further testing (negatives) or (when positive) subjected to additional, confirmatory analyses with other existing Tier-1 assays. This would greatly enhance throughput of chemicals through initial testing, both in terms of resource utilization and timing.

ACKNOWLEDGMENTS

The authors would like to thank the following collaborators: Christopher J. Borgert, Mark Hanson, Alan Hosmer, Werner Kloas, and Glen Van Der Kraak.

DISCLAIMER

The views and statements expressed in this paper are those of the authors alone. The views or statements expressed in this publication do not necessarily represent the views of the organisations to which the authors are affiliated, and those organisations cannot accept any responsibility for such views or statements.

The manuscript has been subjected to review by the National Health and Environmental Effects Research Laboratory and the Office of Chemical Safety and Pollution Prevention and approved for publication. Approval does not signify that the contents reflect the views of the USEPA and mention of trade names or commercial products does not constitute endorsement or recommendation for use by USEPA.

REFERENCES

380 Adams NR. 1995. Detection of the effects of phytoestrogens on sheep and cattle. *J Anim Sci*
381 73:1509-1515.

382 Ankley GT, Bennett RS, Erickson RJ, Hoff DJ, Hornung MW, Johnson RD, Mount DR, Nichols
383 JW, Russom CL, Schmieder PK, Serrrano JA, Tietge JE, Villeneuve DL. 2010. Adverse
384 outcome pathways: A conceptual framework to support ecotoxicology research and risk
385 assessment. *Environ Toxicol Chem* 29:730-741.

386 Ankley, GT, Gray LE. 2013. Cross-species conservation of endocrine pathways: A critical
387 analysis of Tier 1 fish and rat screening assays with 12 model chemicals. Environ.
388 Toxicol. Chem. 32, 1084-1087.

389 Ankley GT, Jensen KM. 2014. A novel framework for interpretation of data from the fish short-
390 term reproduction assay (FSTRA) for the detection of endocrine-disrupting chemicals.
391 *Environ. Toxicol. Chem*. DOI: 10.1002/etc.2708.

392 Borgert CJ, Mihaich EM, Ortego LS, Bentley KS, Holmes CM, Levine SL, Becker RA. 2011a.
393 Hypothesis-driven weight of evidence framework for evaluating data within the U. S.
394 EPA's Endocrine Disruptor Screening Program. *Reg Toxicol Pharmacol* 61:185-191.

395 Borgert CJ, Mihaich EM, Quill TF, Marty MS, Levine SL, Becker RA. 2011b. Evaluation of
396 EPA's Tier 1 Endocrine Screening Battery and recommendations for improving the
397 interpretation of screening results. *Reg Toxicol Pharmacol* 59:387-411.

398 Borgert, CJ, Stuchal LD, Mihaich EM, Becker RA et al., 2014. Relevance weighting of Tier 1
399 endocrine screening endpoints by rank order. *Birth Defects Res (Part B),* 101, 90–113.

400 Kim, HS, Kang TS, Kang IH, Kim TS, Moon HJ, Kim IY, et al. 2005. Validation study of
401 OECD rodent uterotrophic assay for the assessment of estrogenic activity in sprague-

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

dawley immature female rats. *Journal of Toxicology and Environmental Health. Part A* 68(23-24): 2249-62.

Klimisch H-J, Andreae M, Tillmann U. 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Reg Toxicol Pharmacol* 25:1-5.

Schlenk D., Jenkins F. 2013. Endocrine Disruptor Screening Program (EDSP) Tier 1 Screening Assays and Battery Performance. US EPA FIFRA SAP Minutes No. 2013-03. May 21-23, 2013. Washington, DC.

[USEPA] U.S. Environmental Protection Agency. 2009. Endocrine disruptor screening program test guidelines - OCSPP 890.1250: Estrogen Receptor Binding Assay Using Rat Uterine Cytosol (ER-RUC). EPA 740-C-09-005.

[USEPA] U.S. Environmental Protection Agency. 2009. Endocrine Disruptor Screening Program Test Guidelines - OCSPP 890.1300: Estrogen Receptor Transcriptional Activation (Human Cell Line (HeLa-9903). EPA 740-C-09-006.

[USEPA] U.S. Environmental Protection Agency. 2009. Endocrine Disruptor Screening Program Test Guidelines - OCSPP 890.1150: Androgen Receptor Binding (Rat Prostate Cytosol). EPA 640-C-09-003.

[USEPA] U.S. Environmental Protection Agency. 2009. Endocrine Disruptor Screening Program Test Guidelines - OPPTS 890.1550: Steroidogenesis (Human Cell line - H295R). EPA 640-C-09-003.

[USEPA] U.S. Environmental Protection Agency. 2009. Endocrine Disruptor Screening Program Test Guidelines - OPPTS 890.1200: Aromatase (Human Recombinant). EPA 740-C-09-004.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

[USEPA] U.S. Environmental Protection Agency. 2009. Endocrine disruptor screening program test guidelines—OCSPP 890.1600: Uterotrophic assay. EPA 740/C-09-0010. Washington, DC.

[USEPA] U.S. Environmental Protection Agency. 2009. Endocrine disruptor screening program test guidelines—OCSPP 890.1400: Hershberger bioassay. EPA 740/C-09-008. Washington, DC.

[USEPA] U.S. Environmental Protection Agency. 2009. Endocrine disruptor screening program test guidelines— OCSPP 890.1500: Pubertal development and thyroid function in intact juvenile/peripubertal male rats. EPA 740/C-09/012. Washington, DC.

[USEPA] U.S. Environmental Protection Agency. 2009. Endocrine disruptor screening program test guidelines—OCSPP 890.1450: Pubertal development and thyroid function in intact juvenile/peripubertal female rats. EPA 740/C-09/009. Washington, DC.

[USEPA] U.S. Environmental Protection Agency. 2009. Endocrine disruptor screening program test guidelines—OCSPP 890.1350: Fish short-term reproduction assay. EPA 740/C-09/007. Washington, DC.

[USEPA] U.S. Environmental Protection Agency. 2009. Endocrine disruptor screening program test guidelines—OCSPP 890.1100: Amphibian Metamorphosis assay. EPA 740/C-09/002. Washington, DC.

[USEPA] United States Environmental Protection Agency. 2011. Endocrine Disruptor Screening Program, Weight-of-Evidence: Evaluating Results of EDSP Tier 1 Screening to Identify the Need for Tier 2 Testing. Office of Chemical Safety and Pollution Prevention. September 14, 2011.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

446 Van Der Kraak GJ, Hosmer AJ, Hanson ML, Kloas W, Solomon KR. 2014. Effects of atrazine
447 in fish, amphibians, and reptiles: An analysis based on quantitative weight of evidence.
448 *Crit Rev Toxicol* 44(S5):1-66.

449 Walker E, Hernandez AV, Kattan MW. 2008. Meta-analysis: Its strengths and limitations.
450 *Cleveland Clin J Med* 75:431-439.

451

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

Figure 1: Graphical illustration of an adverse outcome pathway. Outcomes at levels 4 and 5 are apical.

Figure 2: Illustration of the combination links in the AOP for reproduction for atrazine in fish, amphibians, and reptiles. The symbols indicate the mean score for relevance and strength and the vertical and horizontal bars 2xSE of the mean score (from data in Van der Kraak et al. 2014)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

459 Table 1. Ability of the Tests in the Tier 1 Battery to Detect Endocrine Activity

Estrogen, Androgen, Thyroid, and Steroidogenesis Pathways	Derivation of Detection Ability
Estrogenic Activity	ER Binding and ERTA Uterotrophic Female Pubertal Fish Short-Term Reproduction Assay
Anti-estrogenic Activity	ER Binding Female Pubertal Fish Short-Term Reproduction Assay
Androgenic Activity	AR Binding Hershberger Male Pubertal Fish Short-Term Reproduction Assay
Anti-androgenic Activity	AR Binding Hershberger Male Pubertal Fish Short-Term Reproduction Assay
Modulation of Steroidogenesis	Steroidogenesis and Aromatase Assays Male and Female Pubertals Fish Short-Term Reproduction Assay
Modulation of Aromatase	Steroidogenesis and Aromatase Assays Female Pubertals Fish Short-Term Reproduction Assay
Altered Hypothalamic-Pituitary Function	Male and Female Pubertals Fish Short-Term Reproduction Assay Amphibian Metamorphosis Assay
Anti-thyroid Activity	Male and Female Pubertals Amphibian Metamorphosis Assay
Thyromimetic Activity	Amphibian Metamorphosis Assay

460

461

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

462 Table 2: Summary of Hypothesis-Based WoE Evaluations for Genistein for the Estrogen Agonist
463 Hypothesis

	Rank 1	Rank 2	Rank 3
Genistein	Vitellogenin in male fish inconsistent (possibly due to route of exposure) [a,b] Uterotrophic assays positive [c]	ERTA activation [d]; observed fish histopath [b], some changes in rat testes [e], some female pubertal changes [e].	ER binding positive; corroborative observations in pubertal endpoints [e]; steroid hormone changes in fish [b].

464 [a] Zhang, L., Khan, I. A., & Foran, C. M. (2002). Characterization of the estrogenic response to genistein in Japanese medaka
465 (*Oryzias latipes*). Comparative Biochemistry and Physiology. Toxicology & Pharmacology : CBP, 132(2), 203-11.

466 [b] Bennetau-Pelissero, C., Breton B, B., Bennetau, B., Corraze, G., Le Menn, F., Davail-Cuisset, B., et al. (2001). Effect of
467 genistein-enriched diets on the endocrine process of gametogenesis and on reproduction efficiency of the rainbow trout
468 *Oncorhynchus mykiss*. General and Comparative Endocrinology, 121(2), 173-87.

469 [c] Kim, H. S., Kang, T. S., Kang, I. H., Kim, T. S., Moon, H. J., Kim, I. Y., et al. (2005). Validation study of OECD rodent
470 uterotrophic assay for the assessment of estrogenic activity in Sprague-Dawley immature female rats. Journal of Toxicology and
471 Environmental Health. Part A, 68(23-24), 2249-62.

472 [d] Ranhotra, H. S. & Teng, C. T. (2005). Assessing the estrogenicity of environmental chemicals with a stably transfected
473 lactoferrin gene promoter reporter in HELA cells. Environmental Toxicology and Pharmacology, 20(1), 42-7.

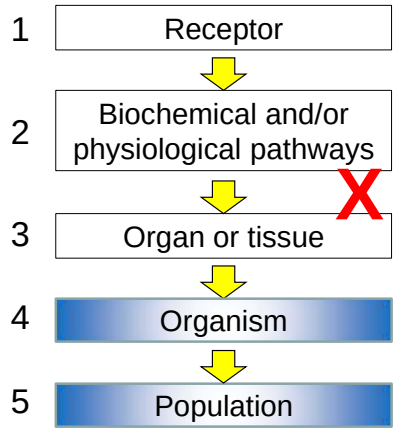
474 [e] Delclos, K. B., Bucci, T. J., Lomax, L. G., Latendresse, J. R., Warbritton, A., Weis, C. C., et al. (2001). Effects of dietary
475 genistein exposure during development on male and female CD (Sprague-Dawley) rats. Reproductive Toxicology (Elmsford,
476 N.Y.), 15(6), 647-63.

477

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

478 Figure 1: Graphical illustration of an adverse outcome pathway. Outcomes at levels 4 and 5 are
479 apical.

480



481

482

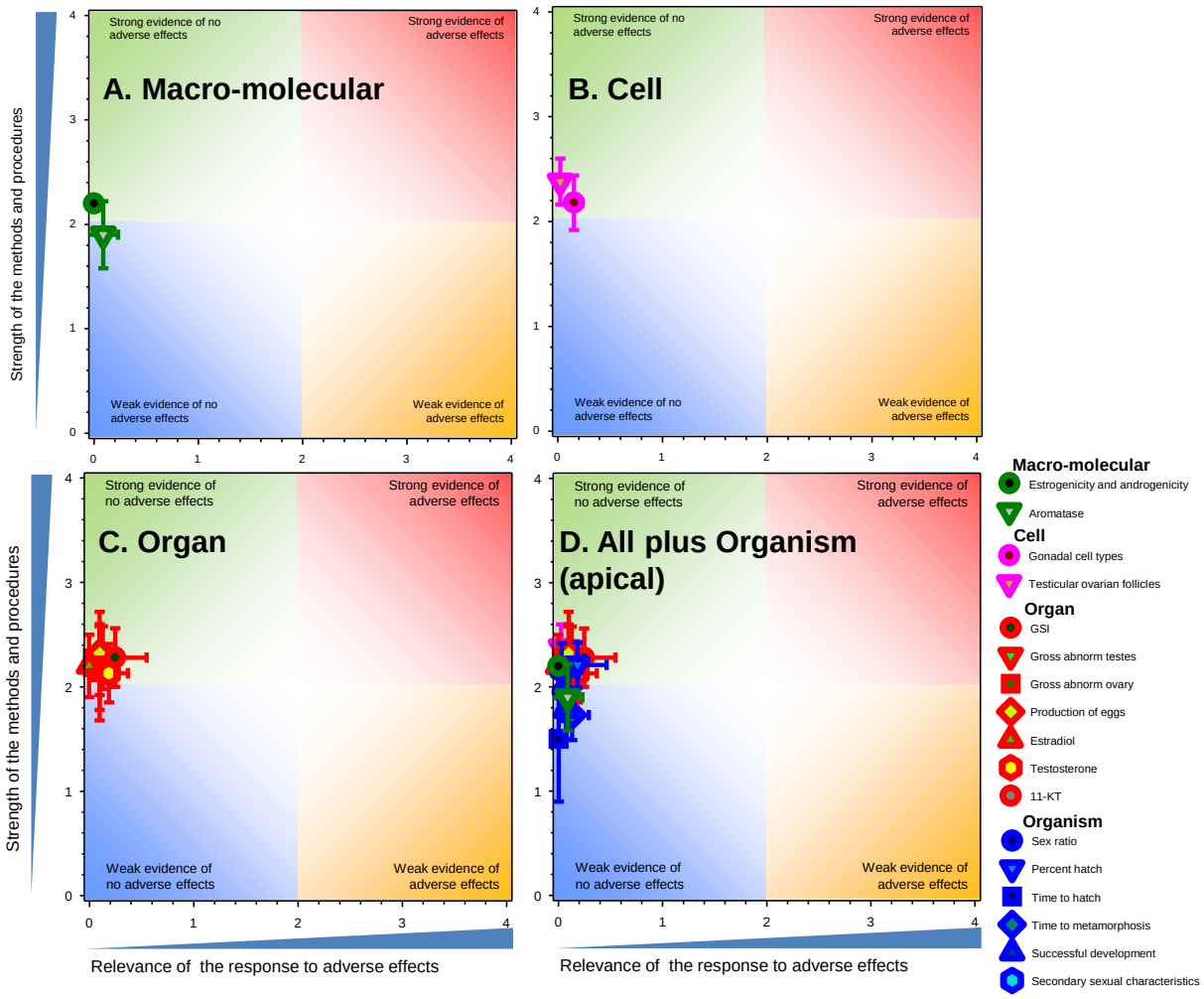
483

484

485

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ:

Figure 2: Illustration of the combination links in the AOP for reproduction for atrazine in fish, amphibians, and reptiles. The symbols indicate the mean score for relevance and strength and the vertical and horizontal bars 2xSE of the mean score (from data in Van der Kraak et al. 2014)



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2596v1 | CC BY 4.0 Open Access | rec: 15 Nov 2016, publ: