



UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
WASHINGTON D.C., 20460

OFFICE OF
CHEMICAL SAFETY AND POLLUTION
PREVENTION

March 2, 2015

MEMORANDUM

SUBJECT: Transmittal of Meeting Minutes of the FIFRA Scientific Advisory Panel Meeting held on "Integrated Endocrine Bioactivity and Exposure-Based Prioritization and Screening"

TO: David Dix, Ph.D., Director
Office of Science Coordination and Policy

FROM: Fred Jenkins, Jr., Ph.D., Designated Federal Official *Fred*
FIFRA Scientific Advisory Panel
Office of Science Coordination and Policy

THRU: Laura Bailey, M.S., Executive Secretary *Laura Bailey*
FIFRA Scientific Advisory Panel
Office of Science Coordination and Policy

Attached, please find the meeting minutes of the FIFRA Scientific Advisory Panel open meeting held in Arlington, VA on December 2-4, 2014. This report addresses a set of scientific issues associated with "Integrated Endocrine Bioactivity and Exposure-Based Prioritization and Screening."

Enclosure

cc:

Jim Jones
Louise Wise
Jack Housenger
William Jordan
Yu-Ting Guilaran
Robert McNally
Donald Brady
Jacqueline Mosby
Jennifer McLain
Dana Vogel
Susan Lewis
Richard Keigwin
Laura Bailey
Tina Bahadori
Rusty Thomas
Jim Cowles
Craig Barber
Steven Knott
Patience Browne
Richard Judson
John Wambaugh
Cathy Milbourn
Linda Strauss
OPP Docket

FIFRA Scientific Advisory Panel Members

James McManaman, Ph.D.
Dana Boyd Barr, Ph.D.
Kenneth Delclos, Ph.D.
David Jett, Ph.D.

FQPA Science Review Board Members

Veronica Berrocal, Ph.D.
Terry Brown, Ph.D.
Robert Denver, Ph.D.
Daniel Doerge, Ph.D.
Edward Perkins, Ph.D.
Thomas Potter, Ph.D.
Catherine Propper, Ph.D.
Daniel Schlenk, Ph.D.
Grant Weller, Ph.D.

FIFRA Scientific Advisory Panel Minutes No.
2015-01

**A Set of Scientific Issues Being Considered
by the
Environmental Protection Agency
Regarding
Integrated Endocrine Bioactivity and
Exposure-Based Prioritization and Screening**

**December 2-4, 2014
FIFRA Scientific Advisory Panel Meeting
Held at the
One Potomac Yard
Arlington, VA**

TABLE OF CONTENTS

PANEL ROSTER	5
LIST OF COMMON ACRONYMS USED	8
INTRODUCTION	9
PUBLIC COMMENTERS	9
OVERALL SUMMARY	10
EXECUTIVE SUMMARY OF PANEL DISCUSSION AND RECOMMENDATIONS	11
DETAILED PANEL DELIBERATIONS AND RESPONSE TO CHARGE	20
REFERENCES	43

NOTICE

The Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA), Scientific Advisory Panel (SAP) is a Federal advisory committee operating in accordance with the Federal Advisory Committee Act and established under the provisions of FIFRA as amended by the Food Quality Protection Act (FQPA) of 1996. The FIFRA SAP provides advice, information, and recommendations to the Agency Administrator on pesticides and pesticide-related issues regarding the impact of regulatory actions on health and the environment. The Panel serves as the primary scientific peer review mechanism of the Environmental Protection Agency (EPA), Office of Pesticide Programs (OPP), and is structured to provide balanced expert assessment of pesticide and pesticide-related matters facing the Agency. FQPA Science Review Board members serve the FIFRA SAP on an *ad hoc* basis to assist in reviews conducted by the FIFRA SAP. The meeting minutes have been written as part of the activities of the FIFRA SAP.

The FIFRA SAP carefully considered all information provided and presented by EPA, as well as information presented by the public. The minutes represent the views and recommendations of the FIFRA SAP and do not necessarily represent the views and policies of the EPA, nor of other agencies in the Executive Branch of the Federal government. Mention of trade names or commercial products does not constitute an endorsement or recommendation for use. The meeting minutes do not create or confer legal rights or impose any legally binding requirements on EPA or any party. The meeting minutes of the December 2-4, 2014 FIFRA SAP meeting held to consider and review scientific issues associated with “Integrated Endocrine Bioactivity and Exposure-Based Prioritization and Screening” were certified by James McManaman, Ph.D., FIFRA SAP Session Chair, and Fred Jenkins, Ph.D., FIFRA SAP Designated Federal Official, on March 2, 2015. The meeting report was reviewed by Laura E. Bailey, M.S., FIFRA SAP Executive Secretary. The minutes are publicly available on the SAP website (<http://www.epa.gov/scipoly/sap/>) under the heading of “Meetings” and in the public e-docket, Docket No. EPA-HQ-OPP-2014-0614, accessible through the docket portal: <http://www.regulations.gov>. Further information about FIFRA SAP reports and activities can be obtained from its website at <http://www.epa.gov/scipoly/sap/>. Interested persons are invited to contact Fred Jenkins, Ph.D., SAP Designated Federal Official, via e-mail at jenkins.fred@epa.gov.

SAP Minutes No. 2015-01

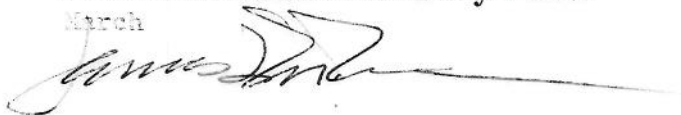
**A Set of Scientific Issues Being Considered by the
Environmental Protection Agency Regarding:**

**Integrated Endocrine Bioactivity and Exposure-Based
Prioritization and Screening**

**December 2-4, 2014
FIFRA Scientific Advisory Panel Meeting
Held at
One Potomac Yard
Arlington, Virginia**

**James L. McManaman, Ph.D.
FIFRA SAP Chair
FIFRA Scientific Advisory Panel**

March



Date: March 2, 2015

**Fred Jenkins, Jr., Ph.D.
Designated Federal Official
FIFRA Scientific Advisory
Panel Staff**



Date: March 2, 2015

PANEL ROSTER

FIFRA SAP Chair

James L. McManaman, Ph.D.

Professor and Chief
Section of Basic Reproductive Sciences
Department of Obstetrics and Gynecology and Physiology
and Biophysics University of Colorado Denver
Aurora, CO

Designated Federal Official

Fred Jenkins, Jr., Ph.D.

US Environmental Protection Agency
Office of Science Coordination and Policy
FIFRA Scientific Advisory Panel
EPA East Building, MC 7201M
1200 Pennsylvania Avenue, NW
Washington, DC 20460
Phone (202) 564-3327 Fax (202) 564-8382 jenkins.fred@epa.gov

FIFRA Scientific Advisory Panel Members

Dana Boyd Barr, Ph.D.

Research Professor
Department of Environmental and Occupational
Health
Rollins School of Public Health
Emory University
Atlanta, GA

Kenneth Delclos, Ph.D.

Research Pharmacologist
Division of Biochemistry Toxicology
National Center for Toxicological Research
US Food and Drug Administration
Jefferson, AR

David Jett, Ph.D.

Director
National Institute of Neurological Disorders and
Stroke
National Institutes of Health
Bethesda, MD

FQPA Science Review Board Members

Veronica Berrocal, Ph.D.

Assistant Professor of Biostatistics
Department of Biostatistics
School of Public Health
University of Michigan
Ann Arbor, MI

Terry Brown, Ph.D.

Professor
Department of Biochemistry and Molecular Biology
Division of Reproductive Biology
The Johns Hopkins Bloomberg School of Public Health
Baltimore, MD

Robert Denver, Ph.D.

Professor and Chair
Department of Molecular, Cellular, and Developmental Biology
The University of Michigan
Ann Arbor, MI

Daniel Doerge, Ph.D.

Food and Drug Administration
National Center for Toxicological Research (NCTR)
Jefferson, AR

Edward Perkins, Ph.D.

Senior Scientist (ST)
Environmental Laboratory
US Army Engineers Research and Development Center
Vicksburg, MS

Thomas Potter, Ph.D.

Research Chemist
US Department of Agriculture Agricultural
Research Station (USDA-ARS)
USDA-ARS, Coastal Plain Experiment Station
Southeast Watershed Research Lab
Tifton, GA

Catherine Propper, Ph.D.

Professor
Biological Services
Northern Arizona University
Flagstaff, Arizona

Daniel Schlenk, Ph.D.

Professor of Aquatic Ecotoxicology and Environmental Toxicology
Department of Environmental Sciences
University of California, Riverside
Riverside, CA

Grant Weller, Ph.D.

Research Scientist
Savvysherpa, Inc.
Minneapolis, MN

LIST OF COMMONLY USED ACRONYMS AND ABBREVIATIONS

AIC:	Akaike Information Criterion
AOP:	Adverse Outcome Pathway
AR:	Androgen receptor
AUC:	Area Under the Curve
BW:	Body weight
CDC:	the United States Centers for Disease Control
EDSP:	the Endocrine Disrupter Screening Program
EDSTAC:	Endocrine Disruptors Screening and Testing Advisory Committee
EEC:	Expected Environmental Concentration
EPA:	the United States Environmental Protection Agency
EPI Suite:	EPA's Estimation Program Interface Suite
ER:	Estrogen receptor
ExpoCast:	EPA's Exposure forecast prioritization research program
FIFRA:	Federal Insecticide, Fungicide, and Rodenticide Act
HEDS:	EPA's Human Exposure Data System
HPDB:	Household Product Database
HPV:	High Production Volume
HT:	High-throughput
HTE:	High-throughput exposure
HTS:	High-throughput screening
HTTK:	High Throughput Toxicokinetics
IBER:	Integrated (RTK) Bioactivity Exposure Ranking
IEMS:	Integrated Exposure Modeling System
IUR:	EPA's Inventory Update Reporting and Chemical Data Reporting list (now CDR)
LEL:	Lowest effect level
LOD:	Limit of detection
NHANES:	National Health and Nutrition Examination Survey
NTP:	the NIH National Toxicology Program November 3, 2014 Page 6 of 111
SAP:	Scientific Advisory Panel
SEEM:	Systematic Empirical Evaluation of Models framework
ToxCast:	EPA's Toxicity foreCast prioritization research program
Tox21:	Toxicology in the 21st Century – the NTP/NCGC/EPA/FDA consortium for chemical hazard HTS
USEtox:	United Nations Environment Program and Society for Environmental Toxicology and Chemistry toxicity model

INTRODUCTION

On December 2-4, 2014 the US EPA Federal Insecticide, Fungicide, and Rodenticide Act Scientific Advisory Panel (FIFRA SAP) met in Arlington, VA to consider and review scientific issues associated with “Integrated Endocrine Bioactivity and Exposure-Based Prioritization and Screening”. The Panel was charged with advising the Agency on its proposed methods of using computational toxicology and exposure tools (i.e., high throughput screening (HTS) assays, and computational models), and other data streams for integrated bioactivity and exposure based prioritization and screening of the universe of Endocrine Disruptor Screening Program (EDSP) chemicals. The Panel provided recommendations to EPA regarding this proposal. Specifically, the Panel addressed the EPA’s charge to them on questions concerning the following three topic areas associated with integrating bioactivity and exposure: 1) estrogen bioactivity, 2) androgen bioactivity, and 3) Integrated Bioactivity Exposure Ranking (IBER). Opening remarks at the meeting were provided by Dr. David Dix, Director of the Office of Science Coordination and Policy. US EPA presentations were provided by the following EPA staff Steven Knott, Patience Browne, Richard Judson, and John Wambaugh.

Presentations were also provided by Warren Casey and Nicole Kleinstreur, respectively staff and affiliate of the US National Toxicology Program’s Interagency Center for the Evaluation of Alternative Toxicological Methods:

PUBLIC COMMENTERS

Oral public comments are listed in the order presented during the meeting:

Patricia Bishop, People on behalf of Ethical Treatment of Animals
Richard A. Becker Ph.D., DABT for the American Chemistry Council
Ellen Mihaich, Ph.D., DABT of the Environmental and Regulatory Resources on behalf of the Endocrine
Sue Yi, Ph.D. of Syngenta on behalf of the Endocrine Policy Forum
Chris Borgert, Ph.D. of Applied Pharmacology and Toxicology on behalf of the Endocrine Policy Forum
Katie Paul Friedman, Ph.D. on behalf of Bayer CropScience and Dow Agrosciences
Sue Marty, Ph.D. D.A.B.T on behalf of Dow Chemical Company
Anna Mazzucco, Ph.D. on behalf of the National Center for Health Research

Written statements were provided by (listed in alphabetical order):

The American Chemistry Council
The Endocrine Policy Forum
People for Ethical Treatment of Animals, Physicians Committee for Responsible Medicine, and the Humane Society of the United States (Jointly written comment)

OVERALL SUMMARY

The Panel was charged with advising the Agency on the following three topic areas associated with integrating bioactivity and exposure: 1) estrogen bioactivity, 2) androgen bioactivity, and 3) Integrated Bioactivity Exposure Ranking (IBER).

In general, the Panel agreed the ER Model for assessing estrogen bioactivity had several strengths. For example, they believed that the ER AUC approach was a computationally, time-resourceful, and insightful approach to determine the estrogenic bioactivity of a chemical. They also noted the AUC ranking had a strong utility for ranking chemicals in association with potential estrogen bioactivity. The Panel concurred that the ER AUC model demonstrated outstanding performance in the characterization of reference chemicals, and in concordance with the selected chemicals for *in vivo* uterotrophic responses.

Although the Panel highlighted several strengths of the ER Model, they also pointed out several limitations and/or weaknesses of the model. For instance, they noted that the models did not incorporate uncertainty or sensitivity analyses. They recommended that the Agency explore the inclusion of such analyses. The Panel also noted that the Agency provide more transparency in describing details about the underpinnings of the model.

In regard to the Agency's efforts to assess androgen bioactivity, the Panel noted that current knowledge suggests that chemicals which impact androgen bioactivity act as relatively weak antagonists rather than agonists. Based upon these largely antagonistic activities of chemicals, they advised that it is critical that the HTS AR bioactivity assay battery include careful assessments and attention to the potential cytotoxic effects of chemicals that may otherwise appear as false positives due to assay interference. This applies to effects generated by the test chemical as well as solvents used in the formulation of chemical doses. The range of chemical structures tested in the assay battery should be expanded to maximize the screening potential. Furthermore, the AR bioactivity battery should include methods to assess the potential effects of chemicals, as well as their metabolites formed by enzymatic conversion in biological systems. The AR bioactivity battery specifically addresses the classical AR nuclear receptor/genomic actions of chemicals, but does not consider other potential non-classical/non-genomic mechanisms that mimic or inhibit androgen bioactivity.

Concerning the Agency's proposed Integrated Bioactivity Exposure Ranking (IBER) approach, the Panel noted that it was rationally developed and laid a foundation for a theoretical basis with the potential to prioritize further EDSP screening of compounds with estrogenic activity. The Panel especially highlighted the practicality of the statistical modeling of the IBER approach. Although the Panel positively remarked about the IBER approach, they cautioned that the approach needed further refinement before it is employed by the Agency's Endocrine Disruption Screening Program. Suggested areas of exploration to further refine the IBER approach included gaining a better understanding of how monitoring data would be to strengthen the approach. The Panel also expressed concern about the limited amount of NHANES exposure data integrated into the IBER model.

EXECUTIVE SUMMARY OF PANEL DISCUSSION AND RECOMMENDATIONS

TOPIC: ESTROGEN BIOACTIVITY

1) EPA's proposed approach for quantifying a chemical's potential estrogen bioactivity is based on a computational model integrating data from 18 high throughput ToxCast assays measuring several endpoints along the estrogen receptor (ER) signaling pathway. The computational model outputs are expressed as area under the curve (AUC) scores for ER agonist (R1) and antagonist (R2) bioactivity. Before routinely using the ER computational model in the Endocrine Disruptor Screening Program (EDSP) framework, EPA is reviewing the scientific strengths and limitations of the ER model described in the White Paper to: i) prioritize chemicals for further EDSP screening and testing based on estimated bioactivity, ii) contribute to the weight of evidence evaluation of a chemical's potential bioactivity, and iii) substitute for specific endpoints in the EDSP Tier 1 battery. Please address the following charge questions relevant to Section 2 of the White Paper and estrogen bioactivity:

a. How clearly has EPA described the computational tools in Section 2.1 (*i.e.*, highthroughput assays and models) used to estimate ER agonist and antagonist bioactivity?

Summary

In general, the EPA has clearly described the computational models used to estimate ER agonist and antagonist bioactivity. However, several areas were noted as providing insufficient information regarding the development and application of the proposed tools. The Panel noted that it would be beneficial if the White Paper provided more information on bioassay performance and if quality control/assurance procedures were made available, either via an addendum or otherwise easily accessible documents. More information is needed regarding the model details such as modified Z-scores, effect of weighting on summary scores, construction of a consensus model, and the penalized least-squares minimization procedure. Additionally, a discussion of the necessity for, and importance of redundancy in terms of orthogonal assays would clarify why so many assays are utilized.

b. What are strengths and limitations of the ER AUC model's ability to identify reference chemicals that include a variety of structures and have a wide range of *in vitro* ER bioactivities?

Summary

Overall, the Panel believed that the ER AUC approach is a computationally time-efficient and intuitive approach to determine the estrogenic bioactivity of a chemical. Its strengths are the use of a combination of assays, its simple and parsimonious mathematical formulation, and its performance on reference chemicals (particularly agonist chemicals). Despite a slighter poorer performance on antagonist chemicals, in general, the ER AUC model provides good or better results than the three assays that it is intended to replace in the existing Tier 1 screening battery.

The limitations of the ER AUC model can be grouped into two general categories: computational/inferential and generalizability. With respect to computational/inferential

limitations, the Panel believed that: (i) efforts should be undertaken to account for uncertainty in the formulation of the statistical model relating receptor signal with assays; and (ii) sensitivity analyses should be conducted to assess how results change under different choices on the model parameters (e.g. the activation threshold or the penalty term in the ER AUC model).

Relative to generalizability, the Panel acknowledged the possibility that the reference chemicals examined to assess the performance of the ER AUC model may not cover all the structural classes that are present in the entire EDSP universe of chemicals.

c. EPA used data from published *in vivo* studies that are methodologically consistent with EDSP Tier 1 guidelines to evaluate concordance between ER AUC model scores of *in vitro* bioactivity, and the *in vivo* uterotrophic response studies (Section 2.2.1). What are strengths and limitations of the curation methods and quality standards used for evaluating published *in vivo* studies?

Summary

The Panel concluded that the methods used to curate and standardize the literature studies for comparing the *in vivo* uterotrophic endpoints to the ER AUC model scores were in most cases sufficient. The ER AUC provided remarkable sensitivity of estimates of uterotrophic responses particularly for low potency compounds with balanced accuracy of 95%, positive predictive value of 97%, and negative predictive value of 92%.

In regard to the limitations of these methods, while the literature evaluations indicated strength of comparisons (*in vivo* to ER AUC) for high and low potency compounds, variability (18/70 chemicals or 26% discordant) was still significant. The Panel agreed with the Agency that this is likely due to differences in design (e.g. sample size) and the number of studies conducted for each chemical.

Overall, with the high degree of variability, and only having one false positive and one false negative out of 42 compounds (Figure 2.12) the ER AUC model is surprisingly accurate. However, additional studies should focus on metabolically activated compounds (e.g. pyrethroids, and methoxychlor) that should be more appropriately detected *in vivo*. In addition, literature curation should also include comparisons with other *in vivo* assays such as the FSTRA and rat pubertal assays.

d. Based on all the data presented in Section 2 on ER AUC model performance including characterization of reference chemicals, and concordance with *in vivo* uterotrophic results, what are strengths and limitations of using the ER AUC Page 2 of 3 model to distinguish and prioritize chemicals based on potential estrogen bioactivity?

Summary

The Panel concurred (assuming linear hER binding and/or activation through nuclear receptors) the AUC model is a relatively simple, but elegant, approach for integrating different bioassay data into a single value that is proportional to the bioactivity or potency of a chemical. The AUC ranking clearly has a strong applicability for ranking chemicals in terms of potential estrogen bioactivity.

Additional strengths of the method include: excellent performance for reference chemicals; consistent outcomes with active and non-active descriptions for List 1, List 2, and the EDSP Universe of compounds; redundancy and power of multiple HTS signals for confirmation of “hits”; strong predictive accuracy of greater than 90%; and strong correlation across potency categories for reference chemicals.

The Panel also noted several limitations including a limited evaluation of variation; a need to use of oral dose equivalents to compare AUC values to the lowest effect level (LEL) in uterotrophic studies; metabolic/biotransformation capacities not included; exclusive AOP focus on nuclear ER without evaluation of alternative receptors/pathways (i.e. GPER); and vague description of potency categories.

Overall, with minor limitations for compounds that require metabolic activation or have targets other than nuclear receptors, the ER AUC appears to be an appropriate tool for chemical prioritization for List 1, List 2 and the EDSP universe compounds.

e. Based on all the data presented in Section 2 on ER AUC model performance including characterization of reference chemicals, concordance with *in vivo* uterotrophic results, and comparison with Tier 1 assay endpoints, what are strengths and limitations of the ER AUC model to contribute to the weight of evidence determination of a chemical’s potential estrogen bioactivity?

Summary

The Panel noted that the ER AUC model had several strengths. They concluded that ER AUC model showed outstanding performance in the characterization of reference chemicals, and in concordance with the selected chemicals for *in vivo* uterotrophic responses (literature and Tier 1 endpoints). The use of the battery of ER AUC fits the AOP for the linear pathway of ER activation, DNA binding of receptor, and trans-activation of the receptor. Use of the AOP paradigm will eventually allow the ability to assess mixtures on combined modes of action or antagonistic pathways once complete molecular pathways have been identified. Use of redundancy and complementary endpoints of ER activation and pathway provides qualitative confirmation of activity.

The Panel also noted several limitations of the ER AUC model which included: an omission of the rainbow trout liver slice data (ER expert system); a general failure to “weigh” modular events in the AOP process; a lack of comparisons between ER AUC data and other *in vivo* FSTRA and Pubertal assays; a lack of consideration for non-monotonicity in ER evaluations; a narrow focus of nuclear ER activation, and lack of metabolic capabilities in the *in vitro* systems.

f. Based on all the data presented in Section 2 on ER AUC model performance including characterization of reference chemicals, concordance with *in vivo* uterotrophic results, and comparison with Tier 1 assay endpoints, what are strengths and limitations of using the ER AUC model to substitute for EDSP Tier 1 ER binding, ER transactivation, or uterotrophic assays for the purpose of characterizing a chemical’s potential estrogen bioactivity?

Summary

The need to rapidly evaluate anthropogenic compounds for estrogen-like activity is critical given that several have been identified as potentially estrogenic. The EPA's efforts to develop a process accommodating this need is commendable. In general, the ER AUC model will capture many, though not all, of the estrogenic compounds in the "Universe of Estrogen Disrupting Compounds." Overall, because both the ER AUC model and the Tier 1 *in vitro* assays capture either nuclear receptor binding and/or transactivation, replacement of the Tier 1 *in vitro* ER endpoints (ER binding and ERTA) with the ER AUC model will likely be a more effective and sensitive measure for the occurrence of estrogenic activity that occurs through nuclear receptor binding and activation.

The Panel found that the data comparing the ER AUC model to the uterotrophic assay were strong for the reference compounds that were clearly estrogenic (high AUC values) or unmistakably not estrogenic (very low AUC values). However, the model outcomes were less straight forward when the ER AUC model for non-reference chemicals was compared to uterotrophic studies where the data were limited or discordant. Furthermore, the data for all of the assays utilizing the List 1 and List 2 battery were below the AUC cut off of 0.1 and were consistently negative for the uterotrophic assay. This finding suggests a very low risk of false negatives in this data set, but was limited by the fact that there were no chemicals with either high or intermediate AUC model values available for functional comparison. A second, but related concern remains regarding the ability of ToxCast *in vitro* methods to assess estrogenic effects that may be impacted by absorption and metabolism or may result through non-nuclear estrogen receptor signaling pathways.

For these reasons, the Panel did not recommend that the uterotrophic assay be substituted by the AUC model at this time. The Panel suggested that the EPA considers: 1) conducting limited uterotrophic and other Tier 1 *in vivo* assay testing, using the original Tier 1 Guidelines (and/or through literature curation), for a limited number of chemicals in the low and middle AUC value range from the "Universe of Chemicals" shown in Fig. 2.16 of the Agency's White Paper. Such testing would then add weight-of-evidence supporting or refuting the hypothesis that the ToxCast suite of assays is predictive of *in vivo* estrogenic responsiveness, and 2) performing similar comparisons of the ER AUC model to the other *in vivo* assays in the Tier 1 estrogenic battery.

TOPIC: ANDROGEN BIOACTIVITY

2) EPA's proposed approach for quantifying a chemical's potential androgen bioactivity is based on a computational model integrating data from nine high throughput ToxCast assays measuring several endpoints along the androgen receptor (AR) signaling pathway. The computational model outputs are expressed as area under the curve (AUC) scores for AR agonist (R1) and antagonist (R2) bioactivity. Before routinely using the AR computational models in the EDSP framework, EPA is reviewing the scientific strengths and limitations of the AR AUC model described in this White Paper to: i) prioritize chemicals for further EDSP screening and testing based on estimated bioactivity, and ii) contribute to the weight of evidence evaluation of a

chemical's potential bioactivity. Please address the following charge questions relevant to Section 3 of the White Paper and androgen bioactivity:

a. How clearly has EPA described the computational tools in Section 3.1 (*i.e.*, high throughput assays and models) used to estimate AR agonist and antagonist bioactivity?

Summary

Results from the battery of assays to estimate AR agonist and antagonist bioactivity of chemicals was presented as preliminary and thereby represents a work in progress. The overall consensus was that the assay battery and computational tools represent a major step forward in defining and quantifying chemical agonist and antagonist activities mediated via the nuclear androgen receptor pathway. In particular, the implementation of high throughput *in vitro* recombinant protein and cell based bioactivity assays that provide an integrated assessment of chemicals affecting the androgen receptor activity pathway was viewed as a significant advancement intended to supplant the narrowly focused, highly variable, labor intensive, and animal-based *in vitro* rat prostate cytosol androgen receptor binding assay in Tier 1 testing. Further refinement of the assay battery was encouraged to optimize the assessment of chemical activities affecting the AR pathway, especially in the context of antagonism of AR bioactivity which defines the action of a majority of chemicals tested to date. Specifically, the assays and computational tools must be sufficiently robust so as to distinguish true AR antagonism from cellular/molecular toxicity. Although the AR test battery is under development, future efforts should be guided by lessons learned from parallel development of the high throughput test battery and computational methods applied to chemical effects on ER agonist and antagonist bioactivity as described in Section 1, Estrogen Bioactivity.

b. What are strengths and limitations of the AR AUC model's ability to identify reference chemicals that include a variety of structures and have a wide range of *in vitro* AR bioactivities?

Summary

The overall strengths of the AR AUC model outweigh the limitations. The use of complementary *in vitro* recombinant protein and cell based assays provide a high throughput modality for the rapid screening of chemicals through an integrated battery of assays for the assessment of AR bioactivity with the benefit of reducing the need for animals. The assays significantly improve the reliability of assessing AR binding (replacing the rat prostate cytosol AR binding assay) and provide new measures of AR transcriptional activity previously not included in the Tier 1 screening. In general, the assays and computational model are designed to distinguish AR agonist and antagonist bioactivities. However, results reported to date cover a rather narrow range of chemical structures and the distribution of calculated values for AR bioactivity fall within a limited range of AR bioactivities. AR AUC values primarily represent known pharmaceuticals with strong agonist activities that appear at the top of the AUC scale or test chemicals with antagonist activities that appear at the bottom of the AUC scale, thus the AUC computational model places these chemicals at disparate ends of the value range. Because chemicals with AR antagonist activity have low values according to the AUC computational model, there is an inherent risk for misinterpretation of cytotoxicity as AR antagonist activity.

c. EPA plans to use data from published *in vivo* studies that are methodologically consistent with EDSP Tier 1 guidelines to evaluate concordance between AR Page 3 of 3 AUC model scores of *in vitro* bioactivity, and the *in vivo* androgenic and antiandrogenic responses (Section 3.2.1). What are strengths and limitations of the planned curation methods and quality standards for evaluating published *in vivo* studies?

Summary

Preliminary data with the assay battery for assessing the *in vitro* AR bioactivity of chemicals correlated well with the *in vivo* Hershberger androgen bioactivity assay data from the literature. This provides initial support that *in vitro* AR bioactivity assays will replicate *in vivo* chemical activities and will be especially beneficial as a rapid, high throughput screen in Tier 1 testing. An obvious limitation of *in vitro* screening assays is the inability to fully appreciate or replicate the multiplicity of biological actions that chemicals produce *in vivo*. Included among these limitations are differences in routes of exposure and dosing, activation or inactivation of chemicals via metabolism, non-genomic androgenic effects and potential off-target effects of chemicals that occur *in vivo* but are not appreciated within *in vitro* assays.

d. Based on the data presented in Section 3 on AR AUC model's performance, what are strengths and limitations of using the AR AUC model to distinguish and prioritize chemicals based on potential androgen bioactivity?

Summary

The AUC model integrates a high throughput battery of assays to probe a linear array of steps in the AR pathway that provides for significant improvement in efficiencies and effectively complements other aspects of the Tier 1 testing program. The test battery incorporates redundancy among different species and different assay methodologies to increase confidence in the values assigned by the AUC model calculations. The downside of assay redundancy lies in the balance that each individual assay contributes to the AUC model and how the so-called "penalty term" is applied to the AUC calculation. The AUC model reliably predicts the agonist and antagonist properties of reference chemicals while minimizing false positives due to overt cellular chemical cytotoxicity. The AUC calculation distinguishes between chemicals with agonist activity at the top of the range and those with antagonist activity at the bottom of the range, however the AUC values are so closely grouped within each category of agonist or antagonist activity that it fails to clearly differentiate between potencies of individual chemicals. The AUC model does not effectively account for the possibility of non-monotonic responses. A broader range of chemicals with structure variation should be explored using the AUC model to provide confidence and demonstrate the robust nature of the model.

TOPIC: INTEGRATED BIOACTIVITY EXPOSURE RANKING (IBER)

3) For Endocrine Disruptor Screening Program (EDSP) chemicals with ToxCast estrogen receptor (ER) and androgen receptor (AR) bioactivity scores (Section 2 and 3), and ExpoCast high throughput toxicokinetics and exposure estimates (Sections 4 and 5), the IBER approach was used to rank chemicals based on the ratio between the bioactivity dose range, and the expected exposure range (Section 6). The IBER approach extends point estimates of bioactivity, toxicokinetics, and exposure for a chemical, to distribution ranges based on uncertainty and population variability. Chemical rankings are based on the ratio of the lower range of the bioactive dose, to the upper range of the exposure estimate. Please address the following charge questions relevant to Section 6 of the White Paper and the IBER approach:

a. How clearly has EPA described the computational tools in Section 6 to develop IBER values, including modeling uncertainty and population variability?

Summary

The Panel noted that the description of the IBER approach to prioritize chemicals was adequately clear. The Panel commended the Agency for developing a metric that aims at capturing the “worst case scenario”, while attempting to account for uncertainty and variability in both chemical bioactivity and population exposure. While this effort is a good starting point, the Panel believed that there are several areas of further development, particularly with respect to modeling and accounting for sources of uncertainty and variability. The Panel encouraged the Agency to develop approaches that account for all the sources of uncertainty and variability and model them jointly rather than via multi-step approaches which fail to propagate the uncertainty with a resulting underestimation of uncertainty. More specific comments and suggestions on approaches to properly account for uncertainty and variability are provided in the Detailed Panel Recommendation portion of this Report.

b. What are strengths and limitations of using the IBER approach to prioritize chemicals for further EDSP screening based on the ratio between the ER bioactivity dose range, and the expected exposure range?

Summary

The Systematic Empirical Evaluation of Models (SEEM) effort was logically developed. It also provided a theoretical framework with the potential to prioritize further EDSP screening of compounds with estrogenic activity through the IBER approach. The statistical models and methods used are complex enough to capture many of the potential sources of variability in bioactivity and exposure. However, they are simple enough to allow for straightforward scientific interpretation, model validation, and most importantly further development. In practice, the applicability and reliability of the IBER approach (for the purpose of identifying really high priority chemicals) depends on how efficiently and appropriately the distribution of bioactivity levels and exposure concentrations were derived. The IBER approach provides a means to take into account population variability via Monte Carlo simulations and in the future through application of the SHEDS model. Therefore, the Panel noted that the approach was

theoretically sound for using data and predictive models to produce estimates with respective uncertainties for both exposure and bioactivity. Using the IBER ratio value is in theory sufficiently “fit-for-purpose” for potentially prioritizing EDSP chemicals.

However, the Panel also expressed several concerns regarding the IBER model that suggest that it will need refinement prior to utilization. There were several limitations that the Panel and the EPA acknowledged and identified. The Panel’s concerns regarding the bioactivity interpretations were already addressed in Question 3a. However, the Panel concurred that there is a need to have better exposure monitoring data available to strengthen the model. The Panel was concerned about the limited amount of NHANES exposure data incorporated into the model. This data set was much less robust in comparison to the data available through ToxCast for bioactivity. The Panel was uneasy with predictions for chemicals for which biomonitoring data were not available or were limited. Not having any or limited data to validate the predictions, the HTE/SEEM models appeared to be more an extrapolation exercise given that the large majority of chemicals do not have measurements available for validation of the model. The Panel commented that the assumptions and parameters used to develop the exposure model had potential problems. The assumptions (such as for example whether or not the pesticide active or inert ingredients is modeled) provide the potential for poorly estimating the exposure data values. Also, the model may need to incorporate more parameters, and/or weight the parameters in place, to help strengthen the model output. The Panel was also concerned that specific human populations such as agricultural workers, chemical formulators and pregnant women, who may have the highest exposure levels for specific compounds were not always taken into account.

The Panel had several recommendations to improve the model. Such as they noted that the inclusion of a Bayesian model formulation that combines the pharmacodynamics and pharmacokinetics together is warranted. Model uncertainty might be addressed by combining predictions from different models together with population variability through Monte Carlo sampling. One Panel member suggested another, and perhaps statistically more appropriate, approach to SEEM, which would be to develop a framework similar to the Collaborative Estrogen Receptor Activity Prediction Project (CERAPP) and combine multiple statistical models together. Another possible alternative approach is to propagate the uncertainty in estimates of exposure and bioactivity forward to produce probability statements about the entire distribution of the ratio of these two values. Lastly the Panel commented that the model should undergo a validation process as what was done with the AUC model approach.

c. What are strengths and limitations of using the IBER approach to prioritize chemicals for further EDSP screening based on the ratio between the AR bioactivity dose range, and the expected exposure range?

Summary

The Panel determined the strengths and weaknesses of the IBER approach to prioritize AR active chemicals for further EDSP screening were largely the same as those identified in their response to Charge Question 3b for ER active compounds. Theoretically, the model is sound for this approach, but it needs refinement followed by validation prior to implementation by EDSP. However, the Panel found that the IBER process utilizing the androgenic assays and data for the exposure modeling were not as far along in their development as they were for the estrogenic IBER. Thus, it was difficult to evaluate whether the approach will work for this set of compounds. In particular, the Panel thought there would be an increased risk of chemicals not becoming prioritized when they should be because of failure of the exposure model component of the IBER approach.

The Panel also noted that when using creatinine adjusted values racial, sex, and age differences in creatinine excretion should be considered. Also, biotransformation products should be considered. Lastly, the Panel emphasized, across all charge questions, that it will be critical to incorporate exposure to complex chemical mixtures in the model to ultimately understand the real exposure risks.

DETAILED PANEL RECOMMENDATIONS

TOPIC: ESTROGEN BIOACTIVITY

1) EPA's proposed approach for quantifying a chemical's potential estrogen bioactivity is based on a computational model integrating data from 18 high throughput ToxCast assays measuring several endpoints along the estrogen receptor (ER) signaling pathway. The computational model outputs are expressed as area under the curve (AUC) scores for ER agonist (R1) and antagonist (R2) bioactivity. Before routinely using the ER computational model in the Endocrine Disruptor Screening Program (EDSP) framework, EPA is reviewing the scientific strengths and limitations of the ER model described in the White Paper to: i) prioritize chemicals for further EDSP screening and testing based on estimated bioactivity, ii) contribute to the weight of evidence evaluation of a chemical's potential bioactivity, and iii) substitute for specific endpoints in the EDSP Tier 1 battery. Please address the following charge questions relevant to Section 2 of the White Paper and estrogen bioactivity:

Question 1a. How clearly has EPA described the computational tools in Section 2.1 (*i.e.*, high-throughput assays and models) used to estimate ER agonist and antagonist bioactivity?

Overall, the White Paper clearly described the conceptual approach that EPA is taking to estimate ER agonist and antagonist activity by assessing 18 assays that measure various responses triggered by interaction with nuclear estrogen receptors. The model described attempts to evaluate activity across a 10,000-fold dose range and take into account interfering effects of cytotoxicity or other non-receptor-mediated triggers of the measured responses. The issue presented for discussion was not whether there are other models that might perform better, but how well this model served its purpose in the prioritization of chemicals that have the potential to interact with nuclear estrogen receptors for further testing. The Agency's use of reference chemicals to evaluate the performance of the model was a very important step. EPA was careful in its description of the process for selecting this set and evaluating the models on it. The EPA should be commended for making ToxCast data and previous model codes publicly available, and the Panel recommended that the R code for the IBER models be made publically available as well. Careful editing of the White Paper would greatly improve its readability and clarity. It was the Panel's understanding that the White Paper is a work in progress; however there were many cases where poor grammar, imprecise wording, and repetition made the document difficult to read.

The Agency stated in their White Paper that "Rather than summarizing or reconsidering the research that has been presented previously, the aim of this document and the subsequent SAP meeting is to request scientific input on EPA's proposed method for applying computational tools". However the Panel believed that a more inclusive background on the relationship of the proposed tools and assays to previously presented work would add greatly to the understanding of the approach and tools. It appeared to the Panel that the Agency was requesting that the computational procedures described in the White Paper be evaluated without consideration of these or other efforts. Presenting the current work in the context of other tools would provide more clarity. As a result, the White paper would be strengthened and more transparent by including a critical discussion of related work. For instance a background or supporting

documentation would compensate for the very limited description of the ToxCast High Throughput Screening assays and their validation in the White Paper. In addition, example it was unclear how the current EPA EDSP ToxCast effort fits into prior Agency work (e.g., Reif et al. 2010; Filer et al. 2014) and other U.S. Tox21 collaborative programs or prior SAP reviews on *in vitro* assays and new high throughput exposure models (e.g., Jan 2013 SAP; July 2014 SAP; Kavlock et al. 2012, Rotroff et al. 2013, Judson et al. 2014; Rotroff et al. 2010, Wetmore et al. 2013, Wetmore et al. 2014; Wetmore et al. 2011, Wambaugh et al. 2013, Wambaugh et al. 2014).

Due to the lack of assay background information, SAP members had several questions regarding the nature, reproducibility and validation of the assays. They were unable to find complete information on these assays via website/database/literature searches to evaluate their strengths and weaknesses. For example, the ‘Assay Providers’ link on the EPA’s ToxCast webpage was nonfunctional at the time of the SAP public meeting. Based on the information presented, it was difficult to tell how these assays were validated, if at all. Since data generated from these 18 assays were used to develop the ER AUC model, the lack of assay information made it difficult to determine if these assays are appropriate to replace Tier 1 assays in the EDSP. Tier 1 EDSP assays were extensively validated prior to evaluation by a prior SAP. It is unclear whether this was done for the ToxCast assays used in the model. If the assays were not validated a rationale should be given in the White Paper. Because of the superficial treatment of the assays, it was not clear that the chosen HTP assays represent the best methodology or that they are indeed state-of-the-art.

It was noted by one Panelist that there were problems with assay interference, and that this varied depending on the type of assay. As a result, the time points chosen for the assays may be too long (8 and 24 hours) to identify direct actions, or may lead to detection of indirect effects on the signaling pathways assayed.

The Panel suggested that the Agency add a summary discussion in the White Paper on the Quality Control/Assurance (QA/QC) procedures used. At a minimum, citations of on-going efforts should be described. For example, some information was available for download from the ToxCast web site. This site provided a summary (as of Oct. 2014) of chemical purity and Molecular Weight validation of chemicals in the ToxCast Universe. There were 12,779 entries with 57% flagged as passed, 36 % as “not determined” (analyses in progress) and 6% as “caution”. Similar metrics on the subset of chemicals examined in *in vitro* estrogen and androgen activity screening programs were not available. Such data would add confidence to results. Other QA/QC metrics that are also critical include results of the replicate analyses with the estrogen and androgen screening matrices. It was noted that area under the curve (AUC) values presented in the text are presented as unit values in most cases. Clarity of these results would be improved by reporting error bounds linked to analysis of blind replicates in the ToxCast batteries.

Although the overall approach was described in the White Paper, the Panel indicated that there were several aspects that were vague. The presentations given by EPA during the public SAP meeting were helpful in clarifying some aspects that were unclear in the written material. For example in the discussion surrounding Equation (2.1), it would have been helpful to have more description regarding the meaning and justification for the use of the modified Z-score, the

penalty term, and the explanation of “discounting” if one receptor was activated at lower doses than others. In particular, an explanation justifying the use of the median and MAD, as opposed to mean and standard deviation, was warranted. Further discussion of the effect of estimation error in the penalized regression model (2.2) would also have been useful. The Panel recommended that the Model goodness-of-fit be investigated.

The Panel also recommended that the description of the penalized least-squares minimization procedure be improved. In particular, the estimation procedure used a method similar to a ridge regression penalty (Hoerl and Kennard, 1970). This methodology appeared to be impromptu, particularly with respect to the turning parameters SR_0 and α . While EPA noted that selected values of these parameters result in sensible results (satisfying the “fit-for-purpose” criterion), the Panel questioned the sensitivity of these results (e.g., R_j estimates, AUC calculations) and the selection of these parameters. The Panel advised the Agency to explore the use of cross-validation to select tuning parameter values. It may also be possible to adapt a hierarchical group LASSO (Zhou and Zhu, 2010) approach for this problem to lend more credibility to the curve-fitting procedure.

The Panel advised that it would be important to add more details to the White Paper regarding the effects of weighting (or the lack thereof) on the integration of different assay results. It is well-known in the statistical research literature that using a weighted combination of multiple models generally improves predictive performance. Thus, the Panel commended EPA for its efforts with the Collaborative Estrogen Receptor Activity Prediction Project (CERAPP) program. However, the description of the consensus model construction warrants more clarity and explanation. There are many different ways to combine the results of multiple predictive models, such as model averaging, decision forests, cross-validation, etc. It is unclear which specific methods were used by the Agency. The White Paper does not disclose which methods the Agency used to combine model predictions. Several Panelists noted that the White Paper description of the use of Z-scores to quantify cytotoxicity was vague (though more clarity was provided in EPA presentations at the public meeting). A specific example for illustration would have been helpful. Additionally, a more thorough explanation of the relationship of dose to AUC calculations would minimize the potential for confusion.

The Panel remarked that it would have also been useful if the Agency provided a more detailed explanation of the potential shortcomings of the estimation methodology (e.g., non-unique solutions), and how these shortcomings might affect “downstream” uses of the estimators (e.g., producing AUC values). In particular, while EPA attempts to quantify modeling uncertainty in IBER in Section 6, it does not appear that all uncertainty is carried forward.

The Panel advised that a section in the White Paper be added to describe and emphasize the importance of redundancy, especially with regards to orthogonal assays even though presentations by EPA during the public SAP meeting clarified many of the Panel’s questions. Although the main goal of the presentations was to assess the utility of this model, it was stated in the White Paper that it would be unlikely that all 18 assays would be available for the assessment of all chemicals in the future. Thus, the Panel noted that some discussion of what subset of assays would be most useful would be helpful.

Question 1b. What are strengths and limitations of the ER AUC model's ability to identify reference chemicals that include a variety of structures and have a wide range of *in vitro* ER bioactivities?

The ER AUC model was found to be a commendable attempt by the EPA to develop a statistical model and a computational tool based on a series of *in vitro* and cell-based assays. The model allows for prediction of potential estrogenic activity of chemicals *in vitro*, thus reducing the need for more expensive animal testing. By analyzing several chemicals in parallel and using a combination of assays, the ER AUC is both computationally and time efficient. It also provides a robust assessment of the *in vitro* estrogenicity of a chemical by allowing detection of technology-specific false positives and negatives.

The ER AUC model has a very simple mathematical formulation based on two main assumptions: (i) the assay signal represents a true, nuclear receptor-mediated process and (ii) there is a linear relationship and no-loss of information in the transmission from the receptor to the assay signal. Despite its simplicity, the results presented in the White Paper indicated that the ER AUC model was able to detect reference chemicals that act as agonists. In particular, results indicated that the model was able to detect weak agonists better than the Tier 1 EDSP assays.

With respect to antagonist reference chemicals, the Panel observed that the model had a slightly less robust capacity of detection when compared to agonists due to the confounding effects of cytotoxicity and cellular stress. Nevertheless, in general, the results produced by the ER AUC model were as good or better than results obtained using the three assays that it is intended to replace in the existing Tier 1 screening battery.

The simplicity, parsimony, and interpretability of the ER AUC model were identified as strengths. Several competing statistical models for predicting estrogenic activities of chemicals based on cell-based assay data might be formulated, assuming different physiological relationships between the cellular signaling pathway and assay signal. It is possible that more complex models provide a slightly better performance when evaluated via the reference chemicals, especially with respect to antagonist chemicals. However, as the priority of the EDSP chemical screening is on estrogen agonists, the Panel noted that it was unlikely that more complex models would yield significantly better results than the ER AUC model. In the presence of an equivalent predictive ability on reference chemicals yielded by various alternative and competing models, the Panel recommended that priority be given to models that are more parsimonious and succinct, as identified by Occam razor's principle (Jefferies and Berger, 1992).

The Panel recognized some limitations with the ER AUC modeling approach, and suggested that the EPA investigate the sensitivity of the ER AUC results to modeling assumptions and estimation methods as well as consider different modeling strategies in an effort to improve the ER AUC predictive performance for antagonist chemicals. The Panel noted that the results of the ER AUC's model for the receptor pathways depend on the pathway signals, which are in turn estimated using a constrained penalized least square approach. It is well reported in the statistical literature that parameter estimates derived using a penalized least squares approach are influenced by the penalization constant. In particular, model parameters "shrink" towards zero as the magnitude of the penalty term changes.

The Panel noted that the penalty term in the ER AUC model has been determined in an *ad-hoc* way and not via cross-validation as is standard practice in this type of problem (Golub et al. 1979; Hastie et al. 2009). As the choice of the penalty term might have an effect on the estimated pathway signal and thus on the ER AUC predictive performance, the Panel encouraged EPA to carry out a sensitivity analysis to determine how dependent the predicted ER AUC values are on the penalty term adopted. The Panel hypothesized that large ER AUC values are likely to be robust to changes in the penalty term, and thus results for high activity chemicals may not change significantly. However, the ER AUC results for moderate or low activity chemicals with ER AUC closer to the activity threshold may be affected by changes in the penalty term.

The adoption of a threshold of 0.1 for ER AUC values to identify highly active chemicals is not clearly defined in the White Paper. The classification of chemicals as true/false positive and, respectively, true/false negatives depends on the threshold value of ER AUC adopted. Given the uncertainty in the estimated receptor signals, which in turn propagated to uncertainty in the ER AUC values, it is plausible that reference chemicals with ER AUC values near the threshold might be incorrectly misclassified (ref. Figure 2.4 in the White Paper). This would change the accuracy of the ER AUC model on the reference chemicals. Although, the Panel recognized that the uncertainty in the estimated receptor signals will likely affect the predicted classification of the moderate/low-activity chemicals which are not the priority chemicals at the moment, the Panel believed that uncertainty in the threshold and consequent classification of chemicals should be acknowledged.

The Panel also noticed that the ER AUC values were derived not by using raw data arising from replicates of assay experiments, but on the best-fitting Hill, Gain-Loss or constant models fitted to the raw data selected using model fitting criteria such as the Akaike Information Criteria (AIC). While Hill or Gain-Loss models have been previously used in the literature, the Panel acknowledged that the model fitting strategy used in the ER AUC model introduces additional uncertainty. The Panel suggested that data arising from each assay replicate be used to derive several curves relating a chemical concentration to an assay signal (each corresponding to a replicate), yielding in turn a range of ER AUC values from which mean/median and confidence intervals could be derived, communicating the uncertainty in the estimated bioactivity of the chemical.

As mentioned above, the ER AUC model was based on the assumption of a linear relationship between assay signal and receptor signal. While the Panel recognized that this is a good starting assumption, there may exist different functional relationships between assay and receptor signals. The Panel encouraged EPA to investigate different types of functional relationships, especially if this endeavor might improve the ER AUC model's predictive performance for antagonist chemicals.

In conclusion, the ER AUC model yielded good predictive results on the reference chemicals, which, constituting the training set, have possibly guided the selection of the assays mostly focused on nuclear estrogen receptors. As discussed in the public presentations, there is uncertainty with how well the chemicals examined to this point sufficiently cover the structural classes that are present in the entire universe of chemicals to be assessed.

Question 1c. EPA used data from published *in vivo* studies that are methodologically consistent with EDSP Tier 1 guidelines to evaluate concordance between ER AUC model scores of *in vitro* bioactivity, and the *in vivo* uterotrophic response studies (Section 2.2.1). What are strengths and limitations of the curation methods and quality standards used for evaluating published *in vivo* studies?

The Panel concluded that the methods used to curate and standardize the literature studies for comparing the *in vivo* uterotrophic endpoints to the ER AUC model scores were in most cases sufficient. The separation of studies between injection and oral dosing strategies to identify potential study design discordance was valid. The separation of studies using immature and adult ovariectomized rodent protocols was also warranted to detect design discrepancies. The Panel noted that future studies should focus on using rats given the insensitivity of mice.

The ER AUC provided remarkable sensitivity of estimates of uterotrophic responses particularly for low potency compounds with balanced accuracy of 95%, positive predictive value of 97%, and negative predictive value of 92%.

In regard to limitations of these methods, while the literature evaluations indicated strength of comparisons (*in vivo* to ER AUC) for high and low potency compounds, variability (18/70 chemicals or 26% discordant) was significant. Although the Panel agreed with the Agency that this is likely due to differences in design and the limited number of studies conducted for each chemical, it could also be due to the following errors, which in some cases may or may not be determined from strict literature evaluation but whenever possible should be evaluated:

1. *Chemical purity and availability in dosing materials.* This item is not a subject specific to literature evaluations, but should be considered in all HTS assays. Recent HTS studies with zebrafish embryos indicated that a contaminant from a flame retardant that was present in less than 1% was responsible for significant adverse effects (Bugel and Tanguay 2015).
2. *Site of injection.* While it was stated that 99% of injections were performed with subcutaneous methods, the site of injection was not available. The site can either be in the neck or abdomen which could likely affect distribution to the portal venous system and first pass metabolism.
3. *Differences in solvent/carrier controls.* It was stated that some studies used different carriers (ethanol vs. saline) and this could lead to variability between studies.
4. *Specifics on strain differences between animals.* While strain differences between animals was stated in the oral presentation; it was not discussed in the Agency's White Paper (i.e. how many "rat" studies used; what strain of rat specimen was used i.e. Sprague Dawley?).
5. *Limited sample sizes.* While the presentation stated that a minimum treatment number was 5, the text indicated control treatment values of 3. This low value is unacceptable and likely contributes to the variance between studies. The original Tier 1 guidelines suggest a minimum sample size of 6 for all treatments. Given the number of treatments (4), a sample size of three in the control group provides very little statistical power.
6. *Systemic toxicity evaluations.* The Panel suggested that it is imperative to maintain strict observations of body weight during exposure as this is a primary indication of overt toxicity which can clearly confound endocrine responses.

The Panel noted that additional methodological limitations included the following:

1. The uterotrophic responses were lacking for LELs between 0.01 and 1 (slide 16/22 Casey Presentation). To have more points on the curve, additional studies that compared uterotrophic responses with LELs are necessary.
2. Figure 2.12-- For Reference compounds there were only 3 values between LELs of 0.2 and 0.001. Similarly with only values provided between higher potency (0.2) and lower potency (0.001) metrics, additional data between these would enhance certainty and may reduce discordancy.
3. Figure 2.15-- For non-reference compounds there was only one value between LELs of 0.02 and 0.001. As mentioned above, additional data points are needed.
4. Concern that literature based curation was not performed on other *in vivo* Tier 1 responses including the FSTRA, rat pubertal.

Overall, with the high degree of variability, and only having one false positive and one false negative out of 42 compounds (Figure 2.12) the ER AUC model was surprisingly accurate. However, additional studies are needed that focus on metabolically activated compounds (e.g. pyrethroids, and methoxychlor which should be more appropriately detected *in vivo*).

Question 1d. Based on all the data presented in Section 2 on ER AUC model performance including characterization of reference chemicals, and concordance with *in vivo* uterotrophic results, what are strengths and limitations of using the ER AUC Page 2 of 3 model to distinguish and prioritize chemicals based on potential estrogen bioactivity?

The Panel concurred (assuming linear hER binding and/or activation through nuclear receptors) that the AUC model is a relatively simple, but elegant, approach for integrating different bioassay data into a single value that is proportional to the bioactivity or potency of a chemical. The AUC ranking clearly has a strong applicability for ranking chemicals in terms of potential estrogen bioactivity. It seems that more information from the bioassay AC50 curves could be utilized to understand uncertainty surrounding the bioactivity AUC value. The Panel noted that specific strengths of the model include:

1. Excellent performance for reference chemicals.
2. Outcomes for Lists 1 and 2 and Universe seemed consistent with active and non-active descriptions.
3. Redundancy and power of multiple HTS signals for confirmation of “hits.”
4. Strong predictive accuracy of greater than 90% and
5. Strong correlation across potency categories for reference chemicals.

The Panel also noted the following limitations:

1. *Loss in Variation in AUCs:* Variation in AUCs between different bioassays was lost with the use of a mean, or median, AUC score for bioactivity from the active assays in the total 18 assays examined. This results in losing measures of uncertainty surrounding the range of active assays for a chemical. Since the potency relationship of the different *in vitro* tests to *in vivo* results may be different, using the lowest AUC for a chemical as a lower bound for bioactivity and the highest as the upper bound may give a better approximation of the *in vitro* bioactivity impact relative to *in vivo* effects. A similar approach is used in Section 6 of

the White Paper where the median and minimum *in vitro* active concentration at the cutoff was calculated.

2. *Comparison of a single number (mean?) AUC value to the Lowest effect Level in an uterotrophic study.* Comparing the mean AUC value to the lowest effect level, appears to be similar to comparing apples to oranges and has a lack of biological relevance between the measures. A more meaningful comparison may be to compare the range of BMD in oral dose equivalents (mg/kg/d) derived from AUCs via biological pathway activation concentrations and HTTK approaches as developed in Section 5 and applied in Section 6 of the Agency's White Paper to BMD values derived from uterotrophic studies (assuming that such studies had appropriate data to calculate a BMD). This may result in a more relevant comparison, as it would include some rudimentary toxicokinetics.
3. *Metabolism not included:* Reference chemicals should include those that are metabolically activated. Metabolic activation can be evaluated using incubations with human and rat s-9 prior to cell treatment. Examples of metabolically active reference chemicals include methoxychlor, permethrin, PCBs, other PAHs. While uterotrophic assays may enhance clearance with greater metabolism, activation to estrogenic metabolites may be significantly important in sensitive developmental windows.
4. *Nuclear receptor focus:* The Panel recommended that the Agency explore alternative receptors/pathways (i.e. GPER) instead of focusing primarily on nuclear receptors.
5. *Selection of potency categories:* The Agency's selection and differentiation of potency categories was vague (i.e. the Panel was unclear about what the term "modest potency" means numerically? (Refer to slide 63/65 in the EPA presentation during the public meeting entitled "Section 2.1:Computational Tools & Models" was much more quantitative.)

Overall the Panel noted that with minor limitations for compounds that require metabolic activation, ER AUC appears to be an appropriate tool for chemical prioritization for List 1, List 2 and EDSP universe compounds.

Question 1e. Based on all the data presented in Section 2 on ER AUC model performance including characterization of reference chemicals, concordance with *in vivo* uterotrophic results, and comparison with Tier 1 assay endpoints, what are strengths and limitations of the ER AUC model to contribute to the weight of evidence determination of a chemical's potential estrogen bioactivity?

The Panel concluded that ER AUC model had several strengths. For example, the ER AUC showed outstanding performance in the characterization of reference chemicals, and concordance with the selected chemicals for *in vivo* uterotrophic responses (literature and Tier 1 endpoints). In addition, use of the battery of ER AUC fits the AOP pathway for the linear pathway of ER activation, and DNA binding of receptor and trans-activation of the receptor. Also use of the AOP paradigm will eventually allow the ability to assess mixtures. This may occur additively on combined modes of action or assess antagonistic chemical pathways once complete molecular pathways have been identified. They Panel also positively noted that use of redundancy as well as complementary endpoints of ER activation and pathway provides qualitative confirmation of activity. This attribute was initially an objective of EDSTAC

(<http://epa.gov/endo/pubs/edspoverview/finalrpt.htm>). This should be maintained unless advances in science dictate otherwise.

In regard to the limitations of the ER AUC model the Panel noted that the rainbow trout liver slice data (ER expert system) should be implemented given the multiple nodes in the AOP process using this tool provides including metabolic activation prior to molecular initiation as well as cellular and protein responses. This tool is not only effective for ecological and wildlife evaluations, it is also effective for overall AOP evaluations for estrogenic activity. In addition, in follow up to the July 2013 FIFRA SAP meeting on “Weight-of-Evidence (WoE): Evaluating Results of EDSP Tier 1 Screening” a more transparent and objective description of the WoE process for prioritization was found necessary. For example, how will various endpoints and potencies be “weighted” in evaluation. Thus, at some point, it will be necessary to quantify modular relationships in the AOP process. ER AUC data should be compared against additional *in vivo* fish short- term reproduction assay (FSTRA) and pubertal assays as these assays have demonstrated significant sensitivity in previous EDSP evaluations (Ankley and Gray 2013). There was also lack of consideration for non-monotonicity in ER evaluations. Multiple studies have observed non-monotonic responses within *in vitro* and *in vivo* systems at low chemical concentrations (Vandenberg 2012 Bergman et al. 2013). This contrasts ToxCast data that had limited examples of non-monotonic responses, and these were exclusively related to cytotoxicity. There was clearly a discrepancy here, and additional evaluations are necessary to determine why it exists. Overall, estrogenic activity is narrowly defined as nuclear ER activation, and it is recommended that additional assays be employed as they become available to assess other estrogenic activity pathways that may not occur through nuclear ER binding or activation. An often stated deficiency with the *in vitro* assays is the lack of metabolic capability to assess the potential for *in vivo* inactivation or activation of the chemical. However, this is for the most part a tolerable deficiency if the question is limited to the potential for nuclear receptor-mediated activity. It would not be tolerable in cases where only a metabolite of the compound was biologically active.

Question 1f. Based on all the data presented in Section 2 on ER AUC model performance including characterization of reference chemicals, concordance with *in vivo* uterotrophic results, and comparison with Tier 1 assay endpoints, what are strengths and limitations of using the ER AUC model to substitute for EDSP Tier 1 ER binding, ER transactivation, or Uterotrophic assays for the purpose of characterizing a chemical’s potential estrogen bioactivity?

Strengths

In general, the ER AUC model will capture many, though not all, of the estrogenic compounds in the “Universe of Estrogen Disrupting Compounds.” Overall both the ER AUC model and the Tier 1 *in vitro* assays capture either nuclear receptor binding and/or transactivation. Thus, replacement of the Tier 1 *in vitro* ER endpoints (ER binding and ERTA) with the ER AUC model will likely be a more effective and sensitive measure for the occurrence of estrogenic activity that occurs through nuclear receptor binding and activation.

Questions regarding whether the ER AUC model can replace the Tier 1 uterotrophic assay are more complicated. The replacement of this *in vivo* assay is desirable from both animal welfare and cost perspectives. The data comparing the ER AUC model to the uterotrophic assay were strong for the reference compounds that were clearly estrogenic (high AUC values) or

unmistakably not estrogenic (very low AUC values), although there were a few false negatives and positives (Fig. 2.12). EPA provided appropriate explanations for these false outcomes. Also, there was some concern expressed in public comments that the compression at the top of the AUC curve may be problematic, the Panel did not have similar concerns. They noted that the compounds at the top of the curve were largely well known estrogen-related reference chemicals, and were validated through both the Tier 1 *in vitro* and *in vivo* assays. The Panel recognized that the concordance of the model at the high and low end of the AUC spectrum was outstanding.

One of the strengths of the ER AUC model was consideration of potency in comparisons to the uterotrophic assay results (Fig. 2.14). If the compounds in the intermediate AUC ranges only experienced active detections at a very high low effect level (LEL) of a compound, the potential for the compound to have exposure and biological relevancy may be questioned. As EPA pointed out, this information can then be taken into account when trying to determine the model's potential for use in place of the uterotrophic assays. Such an approach was provided to the Panel for the curated non-reference chemicals with AUC values below 0.1, but with positive uterotrophic assay results (see Table 2.10 and discussion below), but not for those with positive uterotrophic assay results between 0.1 and 0.4. On the other hand, for the chemicals that provided a negative uterotrophic assay result and AUC model values between 0.1 and 0.4, the data suggested that these may be true negatives since the no effect level (NEL) was high. Results emphasize that taking potency into account may help fine tune the model.

Limitations

The Panel also found limitations to the model. Specifically, model outcomes were less straight forward when the ER AUC model for non-reference chemicals was compared to uterotrophic studies where the data were limited or discordant (Fig. 2.15 in White Paper). In particular, when the AUC was below about 0.4, and even below the AUC cut off of 0.1, there was a mix of active and inactive results. This suggests that the predictability of the model based on the curation approach was limited when the AUC values were lower than 0.4.

Furthermore, the data for all of the assays utilizing the List 1 and List 2 battery were below the AUC cut off of 0.1 and were consistently negative for the uterotrophic assay. This finding has suggested a very low risk of false negatives in this data set, but was limited by the fact that there were no chemicals with either high or intermediate AUC model values available for functional comparison. Specifically, while this data set supported the finding that very low AUC values were predictive of no estrogenic response *in vivo*, they still did not offer insights into the functionality of the model for those compounds in the intermediate ER AUC value range. To summarize this concern statistically, the data presented comparing the ER AUC model to the uterotrophic assay represents "Spearman Karber-like" statistics where doses were symmetrically active and inactive, but data were limited in between.

A related concern was discussed regarding the ability of ToxCast *in vitro* methods to assess estrogenic effects that may be impacted by absorption and metabolism or may result through non-nuclear estrogen receptor signaling pathways. The latter are important physiological processes that are well understood for a few chemicals, but relatively little is understood about many of the other compounds in the "Universe." For this reason, the Panel encouraged the

iteration of this process by comparing ER AUC with the FSTRA and pubertal assays. Furthermore, evaluating concordance of all three *in vivo* assays may provide insights into potential nonnuclear receptor estrogenic action, especially for those compounds with low AUC values (see Table 2.10 and discussion below) when conducted through the ToxCast suite of assays. The Panel recognized that EPA stated that it is currently working towards similar ER AUC model evaluations with the other Tier 1 *in vivo* assays and encouraged this process through comparing ER AUC with the FSTRA and pubertal assays.

For these reasons, the Panel would not recommend that the uterotrophic assay be substituted by the AUC model at this time. The Panel suggested that EPA considers: 1) conducting limited uterotrophic and other Tier 1 *in vivo* assay testing, using the original Tier 1 Guidelines (and/or through literature curation), for a limited number of chemicals in the low and middle AUC value range from the “Universe of Chemicals” shown in Fig. 2.16. Such testing would likely add weight-of-evidence supporting or refuting the hypothesis that the ToxCast suite of assays is predictive of *in vivo* estrogenic responsiveness, and 2) the EPA does similar comparisons of the ER AUC model to the other *in vivo* assays in the Tier 1 estrogenic battery.

TOPIC: ANDROGEN BIOACTIVITY

2) EPA’s proposed approach for quantifying a chemical’s potential androgen bioactivity is based on a computational model integrating data from nine high throughput ToxCast assays measuring several endpoints along the androgen receptor (AR) signaling pathway. The computational model outputs are expressed as area under the curve (AUC) scores for AR agonist (R1) and antagonist (R2) bioactivity. Before routinely using the AR computational models in the EDSP framework, EPA is reviewing the scientific strengths and limitations of the AR AUC model described in this White Paper to: i) prioritize chemicals for further EDSP screening and testing based on estimated bioactivity, and ii) contribute to the weight of evidence evaluation of a chemical’s potential bioactivity. Please address the following charge questions relevant to Section 3 of the White Paper and androgen bioactivity:

Question 2a. How clearly has EPA described the computational tools in Section 3.1 (*i.e.*, high throughput assays and models) used to estimate AR agonist and antagonist bioactivity?

The SAP commended the EPA collaborative team on the work they have performed to provide the information presented in the White Paper describing the early phases of development and evaluation of androgen bioactivity assays relevant to the Tier 1 EDSP (Endocrine Disruptor Screening Program). Importantly, the EPA team provided further explanations during the SAP public meeting that afforded a better understanding of the underlying methodologies and analytical approaches. Specifically, the adaptation and implementation of a battery of *in vitro* HT AR (high throughput androgen receptor) bioactivity pathway assays are a major advancement in the EDSP and represent a response to concepts advocated by previously convened SAPs.

The Panel recommended that the Estrogen Receptor bioactivity discussion of the White Paper be better elucidated similar to the level of clarity expressed during the oral presentations regarding the AR bioactivity assays and model assumptions. They advised that this information be included in a revised version of the White Paper. Testing and model development should be written in the

appropriate historic and scientific context within the White Paper by briefly discussing and summarizing prior efforts that are described in the published literature. A major shortcoming was that there was a limited description of the ToxCast HTS assays in the White Paper. Members of the SAP were unable to find complete information on these assays via website/database/literature searches to evaluate the strengths and weaknesses of the assays. Also, it could not be determined whether or how these assays were validated. Data from eight assays were used to develop the AR AUC model; however, based on the information provided it was difficult to know if these assays were appropriate to replace Tier 1 assays in the EDSP. Other recommendations included a need to identify and discuss quality metrics, including but not limited to purity of chemicals, that were tested in the HTS system and to interrogate reproducibility of results by blind replicate testing of selected compounds with high, medium and low AUC scores. All scores were reported as point values and the uncertainty of these metrics need to be specified. An example of an AUC50 calculation and related discussion of deliberations involving “weighting” or the lack of it for individual tests should be included. The concept of redundancy and/or lack of it in the test battery should also be discussed. Finally it was suggested that factor analysis and/or other computational approaches could be used to determine whether or not it is appropriate to parse the testing tools/assays to a limited set of tests.

Understandably, in contrast to the more extensive evaluation and application of ER bioactivity assays, the AR bioactivity assessments remain a work in progress. The EPA team was encouraged by the Panel to build on the battery of AR bioactivity assays to maximize representation of individual steps in the AR signal transduction pathway as illustrated in the White Paper. This should include AR binding and AR transactivation as in the current repertoire as well as other relevant and definable steps in the pathway. This implies continuing assessment and refinement of the model with particular attention to the contribution of each component assay to the final assignment of the relative AR bioactivity as determined by the AUC model. There was uncertainty among the Panel regarding whether each assay should be assigned an equal value in the AUC model and the calculation of AR bioactivity and what effect, positive or negative, redundant assays (e.g. AR binding assays in multiple species, hAR and cAR) have on the final calculation of AUC and thus on the value for AR bioactivity.

Questions also persisted on how much redundancy in assay design is beneficial and when does redundancy lead to skewing of the data which may result when one step in the AR pathway is over or under-represented in the final AUC calculation. Also, do the current assays capture direct actions of AR bioactivity (e.g. AR binding) as opposed to potential indirect effects, such as in the time-dependent cell-based transcription assays that may measure secondary rather than primary AR-dependent gene regulatory events? The relative absence of information in the White Paper to describe the methodologies and proficiency evaluations for each of the AR bioactivity assays led to some misunderstanding and lack of clarity during the SAP meeting (e.g. the radioligand used in the AR binding assays is the synthetic androgenic ligand, methyltrienolone (R1881) rather than the endogenous androgen ligands, testosterone or dihydrotestosterone). Additional information than is currently available in the White Paper regarding validation of the individual AR bioactivity assays should be readily available. The transparency provided by the EDSP21 dash board is critical to the overall success of the program. It will likely promote confidence of stakeholders and the scientific public, while also providing opportunities for other interested

scientists to offer further input and refinement for the methods, models and approaches employed in these analyses.

Particular attention should be given to issues related to assay interference and to the factors and chemicals that contribute to cytotoxicity and cell stress. Careful assessment of the general properties of solvent and test chemicals in *in vitro* assays should be considered. These factors are critical for the AR bioactivity assays due to the predominance of chemicals, aside from the known pharmaceutical AR agonists, that predominantly express antagonist activity rather than agonist activity.

Whereas the current focus is on the AR nuclear receptor genomic activity pathway, attention should also be given to the development of alternative AR-related assays that do not follow the classical genomic/nuclear receptor pathway.

Question 2b. What are strengths and limitations of the AR AUC model's ability to identify reference chemicals that include a variety of structures and have a wide range of *in vitro* AR bioactivities?

Strengths

Concerning the strengths, the battery of assays described by the Agency are high throughput and probe a linear array of AR dependent bioactivities in a framework that integrates different steps in the AR pathway. The approach efficiently utilizes *in vitro* reconstituted/recombinant protein and cell based assays with a relevant and significant reduction in *in vivo* animal studies. The use of various cell lines with different inherent properties may allow subtle differences in chemical properties to be ascertained by careful assessment of assay performance characteristics and data analyses. The test battery includes assays of AR binding that parallel EDSP Tier 1 studies of AR binding conducted with rat prostate cytosol and these assays are complemented by investigation of additional steps within the AR pathway leading to regulation of target gene transcriptional activity. The assay battery is built on the key aspects of redundancy among different species and the inclusion of different methodologies that employ various protein components and biochemical/molecular detection and quantification tools. In particular, the assay battery demonstrates an ability to distinguish between agonist and antagonist properties of chemicals and attempts to minimize false positives due to overt cellular chemical cytotoxicity.

Limitations

A potential limitation is that the AUC is derived from a network of assays with some redundancy in assay endpoints and some variation in degree of specificity for the AR-dependent pathway that may contribute unequally or disproportionately to the mean calculated AUC value. As presented to the Panel, the AUC value range is narrow and lacks significant magnitude/range for discriminating between AR bioactivity values/scores that assigned to specific chemicals. There was a general lack of clarity regarding how the penalty term is applied to the AUC calculation which may be significant when arriving at a final AUC value for each chemical. The Panel encourages the inclusion of a wider range of chemicals among different structural classes to inform the future studies using these methodologies. The majority of chemicals interacting with

AR have antagonist activity so assays, and AUC values must be able to distinguish between cell toxicity/cell stress and authentic AR antagonism. Efforts should be made to include in the model, assays that address the metabolism of chemicals as well as non-classical AR mechanisms. Further consideration should be given to the ligand of choice for the AR binding assays because methyltrienolone (R1881), although it is a high affinity, non-metabolizable synthetic AR ligand, may possess subtle differences in its binding properties and the AR structural conformations that it induces as compared to the naturally occurring ligands, testosterone and DHT. Specifically, protein-protein interactions of the AR with other regulatory proteins (e.g. coactivators or corepressors) may be sensitive to the reference AR ligand (i.e. R1881, T or DHT). As an example within the current assay battery, interaction of the coactivator, SRC1 with AR is specific to AR bioactivity by agonists, whereas corepressor interactions with AR are more relevant to chemicals with antagonist activity that are prominent among those examined to date.

Other considerations, to date show the relative absence of non-monotonic responses detected in the AR pathway analyses. The Panel recommended that the potential for non-monotonic responses not be excluded from consideration. The current observations may be limited by the predominance of AR antagonism by the chemicals tested and the coincidence of antagonist activity with higher chemical doses whereas receptor-mediated biological effects that are inherently sensitive to lower doses of chemicals may exhibit non-monotonic properties.

Question 2c. EPA plans to use data from published *in vivo* studies that are methodologically consistent with EDSP Tier 1 guidelines to evaluate concordance between AR Page 3 of 3 AUC model scores of *in vitro* bioactivity, and the *in vivo* androgenic and antiandrogenic responses (Section 3.2.1). What are strengths and limitations of the planned curation methods and quality standards for evaluating published *in vivo* studies?

The preliminary data provided by the Agency correlated well with the Hershberger androgen bioactivity assay and thus provides support that *in vitro* AR bioactivity assays will replicate published *in vivo* chemical bioactivities. The curation methods and quality standards were well considered and strong. According to the information provided to the Panel, correlative data will be available on approximately 40 chemicals from articles currently undergoing secondary analyses, and this will provide a significant and relevant but somewhat limited information for evaluation of *in vitro* AR bioactivity data. Both the Hershberger *in vivo* assay and the proposed *in vitro* AR bioactivity assays are designed with parallel objectives to detect either agonist or antagonist activity.

In parallel with the nature and design of many *in vivo* assays, Hershberger *in vivo* assays are more complex than single endpoint *in vitro* assays. The Hershberger assay involves assessments of multiple androgen dependent tissue weights in either the agonist or antagonist mode. The general expectation is that *in vitro* AR bioactivity assays will measure consistent agonist or antagonist activity of chemicals whereas the evaluation of published results from Hershberger assays may show degrees of variation in response among the multiple androgen-responsive target tissues. These variable tissue responses could be related to SARM activity as well as to the diversity of *in vivo* physiological effects generated by a chemical as observed in the Hershberger assay. The numbers of curated studies involving the Hershberger assay may be low due to the relative scarcity of chemicals with defined agonist or antagonist androgenic activity. Because the

design and utilization of the Hershberger Assay dates back several decades, the congruence among *in vivo* study designs for the Hershberger assay may vary significantly due to changes or difference in rat strains, animal care conditions and laboratory environments. *In vivo* test chemical purity, route of administration, dosage, age and strain of rats, may be among the factors that vary significantly between studies and complicate data interpretation for Hershberger assay data published in the literature over the time course of several decades. Systemic toxicities due to chemical exposures may affect results of Hershberger assays and their interpretation. Pertinent study information about animal body weights, feeding behavior, and general tolerance of the dosing schedule will be essential in evaluating Hershberger assay data to distinguish AR dependency from overt toxicity. Metabolism and *in vivo* conversion of parent chemical compounds to active metabolites remains a concern with the current battery of *in vitro* assays. With regard to pesticide metabolites and how AR bioactivity can be affected by metabolic conversion, analyses of vinclozolin and its bioactive metabolites, M1 and M2, are suggested as reference compounds for the ToxCast screen.¹

Question 2d. Based on the data presented in Section 3 on AR AUC model's performance, what are strengths and limitations of using the AR AUC model to distinguish and prioritize chemicals based on potential androgen bioactivity?

Strengths

Regarding the model strengths, the AR AUC model as presented by the Agency displayed considerable value in preliminary evaluations of AR bioactivity with a number of standard chemicals. By making the data available to the scientific public and providing transparency for the computational model(s) used to generate the data, it can be anticipated that additional scientific input will contribute toward the refinement of the model design and its assessments from both within the organization (EPA) as well as by interested external scientists. The AUC model as currently constructed incorporates contributions from a variety of assay designs and endpoint analyses as well as providing filters for cytotoxicity and cell stress. As presented in Section 3, the current data on AR AUC model performance recommend its potential as a tool for prioritization. The data shown in Fig. 3.5, in particular, may serve as a starting point for further validation of the utility of the AR AUC model approach.

Limitations

In regard to the model's limitations due, to the preponderance of chemicals with AR antagonist activities, moving forward with the AR AUC model will continue to require careful discrimination between true AR antagonist activity and effects due to genotoxicity and cell stress. The Panel expressed some concern over the distinction and discrimination of potency defined by the AUC model compared to traditional assays of AR bioactivity. Chemicals tested to date primarily include androgen pharmaceuticals that fall into either the very high (potent agonist) or very low (antagonist) value range according to the AR bioactivity AUC model, with only a small range of AUC values assigned to these chemicals. It should be noted that the data provided by the Agency are preliminary and lack adequate representation of diverse chemical structures. In the opinion of the Panel, the current data set is small and insufficient to adequately validate the approach. Because the current results are themselves preliminary, further

comparisons and correlations with curated *in vivo* responses and data will be necessary. The Panel suggested additional *in vitro* assays, such as HTS for protein structural alterations, be explored to evaluate false positives for AR antagonism that result from chemical effects such as protein denaturation.

TOPIC: INTEGRATED BIOACTIVITY EXPOSURE RANKING (IBER)

3) For Endocrine Disruptor Screening Program (EDSP) chemicals with ToxCast estrogen receptor (ER) and androgen receptor (AR) bioactivity scores (Section 2 and 3), and ExpoCast high throughput toxicokinetics and exposure estimates (Sections 4 and 5), the IBER approach was used to rank chemicals based on the ratio between the bioactivity dose range, and the expected exposure range (Section 6). The IBER approach extends point estimates of bioactivity, toxicokinetics, and exposure for a chemical, to distribution ranges based on uncertainty and population variability. Chemical rankings are based on the ratio of the lower range of the bioactive dose, to the upper range of the exposure estimate. Please address the following charge questions relevant to Section 6 of the White Paper and the IBER approach:

Question 3a. How clearly has EPA described the computational tools in Section 6 to develop IBER values, including modeling uncertainty and population variability?

The Panel agreed that, at a general level, Section 6 illustrated clearly and thoroughly the Agency's IBER approach. In particular, the motivation for and the description of the IBER metric as a "worst-case scenario" were well described in the White Paper. Figure 1.3 also provided a good representation of the fact that the IBER approach accounts for uncertainty and variability in both a chemical bioactivity and population exposure.

With regard to Figure 1.3, Panel members indicated that it would be beneficial to incorporate text in the figure to clearly indicate that IBER values are based on the ratios of the 5-th percentile of the bioactivity distribution over the 95-th percentile of the exposure distribution and not on the overlap of the two distributions. Although Section 6 clearly stated how the Agency derived IBER values, explicitly representing the ratio within Figure 1.3 may alleviate some confusion in the readership.

While the Panel recognized that the IBER approach relies on the computational and statistical methods described in Sections 2-5 for both bioactivity and exposure prediction, the Panel found that as the outputs of those methods are used to construct the IBER ranking, little clarity was provided with respect to which uncertainties and sources of variability are accounted for in the IBER metric. The clarity of the IBER approach would be enhanced by adding more detail on which sources of uncertainty are accounted for, which ones are modeled and how, and which ones are not accounted for and why not. Table 6.1 attempted to provide a very concise description of the approaches used to handle uncertainty and variability, but the Panel indicated that a more in depth elaboration would be helpful.

The Panel made numerous suggestions to the Agency relative to the uncertainty described and assessed in the IBER approach. The suggestions were noted as follows.

It is clear, from the adoption of a Monte Carlo approach that while individual heterogeneity and population variability are accounted for in the pharmacokinetics component of the IBER approach, model uncertainty is currently not addressed (as Table 6.1 indicates). An obvious question is how the Agency intends to account for this. The Panel recommended that the Agency develop strategies to account also for this source of uncertainty.

As Table 6.1 indicated, both model uncertainty and population variability were accounted for in the pharmacodynamics aspect of bioactivity using the “default” method. Although references to Judson (2011) outlining the “default” approach were provided, the Panel noted that providing more detail about this “default” approach would be a valuable addition to the Section.

Generally, the Panel commended the Agency for its efforts in accounting for uncertainty in both the pharmacokinetics and pharmacodynamics aspect of bioactivity; however, the Panel encouraged EPA to formulate a joint Bayesian model for pharmacokinetics and pharmacodynamics that also takes into account uncertainty in the AUC values. This joint model would allow a propagation of the uncertainty from the AUC values, to the pharmacokinetics and pharmacodynamics component of bioactivity providing a more faithful representation of model uncertainty in bioactivity. Finally, a Monte-Carlo approach, as currently used and described in Section 5, would account for population variability.

With respect to the exposure prediction component of the IBER approach, the Panel stated that description of the methods utilized to derive exposure are too general and did not provide enough detailed information. For example, more details on the HTE and SEEM models could be provided to address questions such as: What is the equation that links parent chemical concentration to measured chemical biomass concentration? What prior assumptions are utilized to link parent chemical concentration to concentration of chemical in plasma? What routes of exposure and degrees of biotransformation are considered and used in the reverse prediction model and why? How are creatinine adjusted values used in SEEM? Have other approaches such as urinary excretion rate (UER) considered since these data (or data needed to calculate UER) are now collected in NHANES? Are exposure mixtures considered and/or will they? This information should be provided to enable more transparency on the Agency’s approach.

Although Table 6.1 indicated that model uncertainty had been accounted for in the exposure prediction component, the Panel had concerns that it was not properly accounted for. In particular, the Panel found Figure 4.4 troubling as it indicates that the same level of uncertainty in exposure is predicted for those chemicals for which information exists through NHANES and for those for which no NHANES data are available. Conceptually, this result implies that adding NHANES exposure adds no value to the ExpoCast estimates. The Panel suspected that the unexpected results presented in Figure 4.4 were due to the way the predictions from the HTE are used in the SEEM model, with the chemical parent concentration predicted by the Bayesian model for HTE probably incorporated in the SEEM only through their posterior summary. A better way to propagate uncertainty from the HTE to the SEEM predictions would be to combine the two models in a single Bayesian model. This would allow the predictions for those chemicals for which no data are available to have greater uncertainty.

As Table 6.1 also indicated, population variability was not accounted for in the exposure component. However, as discussed during the public meeting, the Agency is planning to address this by using SHEDS-HT in the next phase. As this is a major and important innovation that would allow the Agency to account for racial, age and sex differences in chemical excretion (e.g. creatinine), the Panel recommended inclusion of this discussion in the White Paper.

While the Panel commended the Agency for its effort to take into consideration uncertainty in the IBER approach, it would also like to point out that there are several sources of uncertainties and variability at play: “parameter estimation uncertainty” (which can be accounted and communicated via confidence intervals), “model uncertainty” (that is, the uncertainty in the formulations of the statistical models used to make predictions, e.g. the linear model in SEEM, the HTE model, the ER AUC model, etc.) and “population variability”. Currently, the IBER ranking approach is accounting, in part, for estimation uncertainty and population variability, but not model uncertainty. The predictions of both bioactivity and exposure are derived conditionally on the adopted model formulation, e.g. the ER AUC model and the HTE/SEEM model. The Panel encouraged the Agency to clarify that the model uncertainty accounted for in IBER is in reality “parameter/estimate uncertainty.”

Model uncertainty could be addressed in each aspect of the IBER approach (including both the bioactivity and exposure component) by considering competing statistical models simultaneously and combining their predictions together using approaches such as super learner (Van der Laan et al. 2007), Bayesian model averaging (Hoeting et al., 1999), or boosting (Friedman et al., 2000; Freund, 2001; Freund and Schapire; 1997) similar to the efforts the EPA is undertaking through Collaborative Estrogen Receptor Activity Prediction Project (CERAPP).

In summary, the Panel recognized that the IBER approach is a scientifically sound approach to provide a single metric to prioritize EDSP chemicals for further study. The Panel also indicated that increased transparency is needed regarding how the ranked chemicals will be tested and whether IBER values will be re-evaluated as the IBER approach undergoes refinement, e.g. will the IBER be performed on a specified timeline, or will it be performed as new data become available?

Finally, the Panel advised the EPA to provide further discussion about the confidence in the performance of the IBER inputs for the extreme cases (e.g., the obviously bioactive chemicals or very low-exposure chemicals). The Agency suggested during the public SAP meeting that these cases are not of high interest because there is a high scientific confidence in the priority levels of these chemicals (very high or very low). Given this position, the Panel noted that the practical impacts (for ranking purposes) of issues such as the “compression” at the upper end of the forecast distribution of bioactivity (discussed during the public SAP meeting) likely become negligible.

Question 3b. What are strengths and limitations of using the IBER approach to prioritize chemicals for further EDSP screening based on the ratio between the ER bioactivity dose range, and the expected exposure range?

The Systematic Empirical Evaluation of Models (SEEM) effort was logically developed and has provided a theoretical framework with potential to prioritize further EDSP screening of compounds with estrogenic activity through the IBER approach. The problem the EPA is trying to address is extremely complex, especially given the lack of information on a vast majority of the chemicals in the EDSP, as well as the large number of chemicals to be considered in the “EDSP Universe.” Furthermore, the EPA’s utilization of uncertainty and variation to identify the lower bioactivity ranges and higher exposure ranges experienced by the general population suggests that the prioritization will be conservative for inclusion of compounds that may have endocrine disrupting capacity. The statistical models and methods used were complex enough to capture many of the potential sources of variability in bioactivity and exposure, yet simple enough to allow for straightforward scientific interpretation, model checking, and most importantly, further development. In practice, the applicability and reliability of the IBER approach at identifying high priority chemicals depends on how well and appropriately the distribution of bioactivity and exposure are derived. Both components of IBER (bioactivity and exposure level determinations) require the EPA to thoroughly account for all sources of uncertainty in the bioactivity and exposure aspect of IBER including model parameter uncertainty, model uncertainty and population variability. The IBER approach accounts for population variability via Monte Carlo simulations and, in the future, through application of the SHEDS model. Therefore, the Panel found that in principal, the proposed approach to use data and predictive models to produce estimates of both exposure and bioactivity, as well as their respective uncertainties, is sound. In summary, using the IBER ratio to prioritize EDSP chemicals likely meets the “fit-for-purpose” criteria proposed by the Agency. However, the Panel also expressed the following concerns regarding the IBER model that suggest that it will need refinement prior to utilization.

The Panel’s concerns regarding the bioactivity interpretations were already addressed in Question 1, but there was consensus that there is a need to have better exposure monitoring data available to strengthen model outcomes. The Panel was concerned about the limited amount of NHANES-mined exposure data incorporated into the model. This data set appears to be much less robust in comparison to the data available through ToxCast for bioactivity.

The Panel understood that the universe of ESDP chemicals is large and that there is a need to derive predictions of exposure so that risks can be assessed. However, the Panel was not confident with predictions for chemicals for which biomonitoring data is not available or was limited. Not having any or limited data to validate the predictions, the HTE/SEEM models appeared to be more an extrapolation exercise given that the large majority of chemicals do not have measurements available for validation of the model. The model’s uncertainty led to a number of questions by Panel members regarding the functionality of the model: is the relationship in the SEEM prediction model deriving the parent chemical exposure really linear? Is there a threshold effect? Are the indicators in the model the SEEM model clearly defined?

One place where uncertainty exists in the model that should be perhaps be addressed within the model framework, is the fact that the uncertainty in the SEEM predictions should be different for chemicals for which the HTE/SEEM predictions utilizes some information from biomonitoring data or for those which do not. Furthermore, it appears that some effort should be devoted to determine whether biomonitoring data for other chemicals beyond the NHANES data set are

available in the literature. Additional data would likely improve confidence in the SEEM predictions.

The Panel was also concerned that specific human populations who may have the highest exposure levels for specific compounds were not taken into account. In particular it was recommended that workplace exposures should be addressed. In the case of pesticides, agricultural workers likely have much higher exposure risks. Numerous studies have shown that formulators, farm workers and their families may be exposed to higher levels of pesticides (for example, McKone et al., 2007; McKelvey et al., 2013; Johnson et al., 2014; Morgan et al., 2014). As pointed out by the public commenters, worker/applicator exposure is routinely addressed in deliberations involving pesticide registration, and therefore, this group is of particular concern. Curating data, such as those referenced above, could be used to calibrate the model and may be particularly important in identifying the high end of the exposure distribution range. A similar issue came up with specific groups who may be more vulnerable to lower exposure levels. It was recommended that efforts be made to define exposures that are characteristic of pregnant and lactating women and infants. It is widely recognized that these populations may be at highest risk due to exposure of endocrine active chemicals. One approach may be to create subclasses with use classifications within the models that delineate chemicals of concern.

The assumptions and parameters used to develop the exposure model had potential problems. The assumptions (such as pesticide active or not, for example) provide the potential for poorly estimating exposure data values. Also, the model may need to incorporate more parameters, and/or weight the parameters in place, to help strengthen the model output. In this case, the triclosan example demonstrated that wrong assumptions in only one parameter can lead to large discrepancies that negatively impact prioritization decisions. Along these lines, within the White Paper, exposure computations involved classifying chemicals according to use patterns and weighting production volumes on this basis. Pesticides were given a low ranking since most are used in agricultural settings where exposure potential to the general population is limited. However, there are many pesticides used in and around homes, schools, etc. where “near-field” exposures are likely. In this further analysis of study data, the Panel quantified the potential exposures and intake doses of 129 preschool children, ages 20 to 66 months, to 16 pesticides (eight organochlorines, two organophosphates, three pyrethroids, and three acid herbicides). In conclusion, these children were likely exposed daily to several pesticides from several sources and routes at their homes and daycares.

It was strongly recommended at the July of 2014 SAP that the pesticide group, and other groupings as necessary, be further subdivided and weighted by near-field and far-field exposure potential. The Panel supported this recommendation. Again, and as noted by the Agency during the public SAP meeting, such an approach would likely have improved computational results for triclosan that were described at the Panel meeting.

With regard to the exposure model, the FIFRA SAP that convened in July of 2014 also felt that the assumption of a log-normal distribution in computations for values less than the detection limit may not be appropriate. Alternate approaches were identified including the possibility that “range-finding” calculations could be used by assuming all LOD were zero or equivalent to the

LOD of detection in separate computations. The later approach would provide more confidence in results.

The Panel believed that the IBER approach in its current formulation did not appropriately account for model uncertainty. The IBER approach used the ER AUC model to determine the *in vitro* ER activity of a chemical, as well as a particular pharmacodynamics and pharmacokinetics models to estimate exposure data. The exposure model assumed linearity of the toxicokinetics. This may not be easily defended for many compounds. Given that there are both kidney and liver associated parameters involved in the toxicokinetics, the Panel questioned the assumption of linearity in the model. Furthermore, the assumptions of the model in general did not take age and the associated shifts in toxicokinetics that may occur into account. Another limitation of the approach was that the derived IBER quantity is a point estimate, whereas a full distribution or confidence interval may be more informative. One aspect of uncertainty addressed by the EPA was population variability, and the Panel suggested that 1) model parameter uncertainty, 2) overall model uncertainty and 3) different modeling formulations could be considered in model development.

The Panel found it encouraging that the IBER model appropriately classified the pharmaceuticals. However, when these results were removed, the classification of other chemicals based on the distance between exposure and toxic equivalents is not as convincing. Detailed assessments of these chemicals within the IBER context need to be conducted to determine if the IBER system can be used in classifications schemes that may be used to prioritize chemicals for further EDSP testing.

This model is predicated on human exposure data. In the future, the Panel suggested that the lessons learned regarding the application of the process to human health risk be applied to the comparative Tier 1 assays and potentially incorporated with wildlife exposure data.

To summarize, overall the Panel was appreciative that the EPA recognized that uncertainty and variability should be incorporated into the model for the IBER approach prior to launching the platform for prioritization. The Panel emphasized that these issues need to be further evaluated and especially fine-tuned by including the uncertainty and variation that may occur to the most vulnerable groups.

One suggestion that the Panel proposed was to include a Bayesian model formulation that combines the pharmacodynamics and pharmacokinetics together. This approach could propagate the uncertainty through the two components and address model parameter uncertainty. Model uncertainty may be addressed by combining predictions from different models together and population variability through Monte Carlo sampling. Given that there exists many potential models and no model is perfect, nevertheless, one panel member suggested another, perhaps statistically more appropriate, approach to SEEM which would be to develop a framework similar to CERAPP as mentioned in the White Paper. This model combines multiple statistical models together. Some methods that could be used to create this combined model include Bayesian model averaging, super learner, random forest, boosting, etc.). There was some discussion about how to construct the IBER metric. As derived, it is a ratio of extreme points of the bioactivity (lower confidence interval bound) and exposure (upper confidence bound), but

measures of central tendency were also discussed. The Panel noted that a more encompassing approach is to propagate the uncertainty in estimates of exposure and bioactivity forward to produce probability statements about the entire distribution of the ratio of these two values. The model as it relates to the EDSP would need the next steps of the validation process as has been done with the AUC model approach including: choosing certain reference compounds, evaluating them through the ToxCast and ExpoCast, getting a ratio value, and comparing the results to literature and Tier 1 *in vivo* output for compounds that are suggestive of ER action in each of the prioritization categories. In theory, the highest priority compounds identified through the IBER model approach should provide the highest estrogenic outcomes from the *in vivo* assays.

Question 3c. What are strengths and limitations of using the IBER approach to prioritize chemicals for further EDSP screening based on the ratio between the AR bioactivity dose range, and the expected exposure range?

The Panel determined that strengths of the IBER approach to prioritize chemicals for further EDSP screening were largely the same as those identified in 3b for ER active compounds. Theoretically, the model is sound, but it needs refinement followed by validation prior to use. Regardless of imperfections in the IBER approach that make the data appear less robust, the purpose of this approach is for prioritization of testing, not for making regulatory decisions. One Panel member thought as such, it would be unwise to invest a great deal of money into refining this approach if less than 50% improvement were to be made in the estimates.

The Panel believed the current versions of the models for both androgens and estrogens demonstrate the potential for predictability only for those compounds that are very strongly androgenic or estrogenic. Those in the intermediate range of biological activity will need validation through all stages of model development. The Panel found the process of model development occurring appropriately. Thus, as reevaluations continue to be performed appropriate chemicals will be captured for prioritization as needed.

Specifically, regarding the IBER process for the AR evaluations, the Panel found that the androgenic assays and data for the exposure modeling were not as far along well developed making it more difficult to evaluate whether the approach will work for this set of compounds. In particular, the Panel thought there would be an increased risk of chemicals not becoming prioritized when they should be because of the failure of the exposure model side of the IBER approach.

The Panel indicated that the use of biomonitoring data in the SEEM approach was a strength, however, the Agency should keep in mind the limitations of NHANES data and seek to obtain data from other studies/resources that specifically: (1) target susceptible populations like pregnant women, fetuses, children and the elderly and (2) populations at higher exposure risk like pesticide formulators and agricultural workers. Also, NHANES does a good job at capturing baseline exposures (e.g., dietary or continual exposures) but does a poor job at capturing episodic exposures, especially those during critical windows of development. In addition, NHANES sampling logistics prevent the collection of data in less dense population areas (some areas where many chemicals are used widely) and do not necessarily capture seasonal exposures (note: these

concerns were also applicable to the questions regarding the IBER approach for ER bioactivity). Also, the long times spent in the MEC units collecting samples and biometry and anthropometric data may allow for the elimination of short lived chemicals (or create MEC specific exposures) thus underestimating exposures. Thus, inclusion of studies without these limitations, although they likely have smaller number of participants were considered imperative to get accurate picture of US chemical exposures.

Other concerns regarding the use of the available exposure data were expressed by the Panel that apply to both the ER and AR models. For example, when using creatinine adjusted values, racial, sex and age differences in creatinine excretion should be considered. A better approach may be to use urinary excretion rate. Also, biotransformation products should be considered. For example, the data showed unremarkable results for several phthalates, but many reports suggest that the monoesters (simple hydrolytic metabolites) are the bioactive form. As the assay is now performed, such bioactivity would not be demonstrated and may be overlooked in the IBER process/ranking. One Panel member thought this issue to be the biggest weakness in the approach.

Lastly, the Panel, across all charge questions, emphasized that it will be critical to incorporate exposure to complex chemical mixtures in the model to ultimately understand the real exposure risks. Cumulative and mixture were not exposures addressed. Since these chemicals are all undergoing the same bioactivity assays, and we know that chemical exposures do not occur in isolation, cumulative exposures to multiple chemicals should be considered in the ranking process. For example, if BPA and DEHP both have a low IBER ratio, they may not be ranked high in the priority process. But if people are commonly exposed to both, which they are, then this may increase the likelihood that pharmacologically relevant effects could be seen by these exposures.

References

- Eskenazi, B, Bradman, A., Gladstone, E., Jaramillo, S., Birch, K., Holland, N. (2003) CHAMACOS, A Longitudinal Birth Cohort Study: Lessons from the Fields *Journal of Children's Health*, , Vol. 1, No. 1 : Pages 3-27
- Freund, Y. (2001). An adaptive version of the boost by majority algorithm. *Machine Learning* 43, 293–318.
- Friedman, J. Hastie, T. and Tibshirani R. (2000). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics* 28, 337-407.
- Giffe T. Johnson, Steve Morris, James D. McCluskey, Ping Xu, Raymond D. Harbison (2014). Pesticide Biomonitoring in Florida Agricultural Workers. *Occupational Diseases and Environmental Medicine*, , 2, 30-38.
- Freund, Y. and Schapire, R.E. (1997). A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences* 55, 119–139.
- Hoeting, J.A. Madigan, D. Raftery, A.E. and Volinsky, C.T. (1999). Bayesian model averaging: a tutorial. *Statistical Science* 14, 382-417.
- Jefferys, W. and Berger, (1992). Ockham's razor and Bayesian analysis. , `American Scientist', (80), 64-72.
- Judson, R.S., Houck, K., Martin, M., Knudsen, T., Thomas, R.S., Sipes, N., Shah, I., Wambaugh, J. and Crofton, K. (2014). "In vitro and modelling approaches to risk assessment from the U.S. Environmental Protection Agency ToxCast program." *Basic & Clinical Pharmacology & Toxicology* 115(1): 69-76
- Kelce WR, Monosson E, Gamcsik MP, Laws SC, Gray LE Jr. (1994) Environmental hormone disruptors: evidence that vinclozolin developmental toxicity is mediated by antiandrogenic metabolites. *Toxicol. Appl. Pharmacol.* Jun;126(2):276-85. PMID: 8209380
- M.J. Van der Laan, E.C. Polley, and A.E. Hubbard (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology* 6(25), Article 25.
- Mckone, T, Castorina, R., Arnly M., Yukuwarbara, Eskenazi, Bradma A, N. (2007). Merging Models and Biomonitoring to Characterize Sources and Pathways of Human Exposure to Organophosphorus Pesticides in the Salinas Valley of California. *Environ. Sci. Technol.*, 41, 3233-3240.

Morgan M., Wilson, Nancy K. and Chuang, J. C. (2014). Exposures of 129 Preschool Children to Organochlorines, Organophosphates, Pyrethroids, and Acid Herbicides at Their Homes and Daycares in North Carolina. *Int. J. Environ. Res. Public Health*, 11, 3743-3764.

Reif, D.M., Martin, M.T, Tan, S.W., Houck, K.A., Judson, R.S., Richard, A.M., Knudsen, T.B., Dix, D.J. and Kavlock, R.J. (2010). "Endocrine profiling and prioritization of environmental chemicals using ToxCast data." *Environmental Health Perspectives* 118(12): 1714-20.

Rotroff, D.M., Dix, D.J., Houck, K.A., Kavlock, R.J., Knudsen, T.B., Martin, M.T., Reif, D.M., Richard, A.M., Sipes, N.S., Abassi, Y.A., Jin, C., Stampfl, M. and Judson, R.S. (2013). "Real-Time Growth Kinetics Measuring Hormone Mimicry for ToxCast Chemicals in T-47D Human Ductal Carcinoma Cells." *Chemical Research in Toxicology*.

Rotroff, D.M., Wetmore, B.A., Dix, D.J., Ferguson, S.S., Clewell, H.J., Houck, K.A., Lecluyse, E.L., Andersen, M.E., Judson, R.S., Smith, C.M., Sochaski, M.A, Kavlock, R.J., Boellmann, F., Martin, M.T., Reif, D.M., Wambaugh, J.F. and Thomas, R.S. (2010). "Incorporating human dosimetry and exposure into high-throughput *in vitro* toxicity screening." *Toxicological Sciences* 117(2): 348-358

Wambaugh, J.F., Wang, A., Dionisio, K.L., Frame, A., Egeghy, P., Judson, R., Setzer, R.W. (2014). "High Throughput Heuristics for Prioritizing Human Exposure to Environmental Chemicals." *Environmental Science & Technology* 2014 Oct 24.

Wendy McKelvey, Bryan Jacobson, Daniel Kass, Dana Boyd Barr, Mark Davis, Antonia M. Calafat, and Kenneth M. Aldous. (2013). Population-Based Biomonitoring of Exposure to Organophosphate and Pyrethroid Pesticides in New York City. *Environmental Health Perspectives* • volume 121 | number 11-12 | November-December. 1349-1356.

Wetmore, B.A., Wambaugh, J.F., Ferguson, S.S., Li, L., Clewell, H.J., Judson, R.S., Freeman, K., Bao, W., Sochaski, M.A., Chu, T.M., Black, M.B., Healy, E., Allen, B., Andersen, M.E., Wolfinger, R.D. and Thomas, R.S. (2013). "Relative Impact of Incorporating Pharmacokinetics on Predicting In vivo Hazard and Mode of Action from High-Throughput In vitro Toxicity Assays." *Toxicological Sciences* 132(2): 327-346.

William H. Jefferys, James O. Berger (1992). "Ockham's razor and Bayesian analysis", *American Scientist* 80, 64-72