Chemical Hazard Prediction and Hypothesis Testing Using Quantitative Adverse Outcome Pathways

Edward J. Perkins [1], Kalyan Gayen [2], Jason E. Shoemaker [3], Philipp Antczak [4], Lyle Burgoon [1], Francesco Falciani [4], Steve Gutsell [5], Geoff Hodges [5], Aude Kienzler [6], Dries Knapen [7], Mary McBride [8], Catherine Willett [9], Francis J. Doyle III [10] and Natàlia Garcia-Reyero [1]

[1]US Army Engineer Research and Development Center, Vicksburg, MS, USA; [2]Department of Chemical Engineering, National Institute of Technology Agartala, Barjala, Jirania, West Tripura, Tripura, India; [3]Department of Chemical and Petroleum Engineering, Swanson School of Engineering, University of Pittsburgh, PA, USA; [4]University of Liverpool, Liverpool, Merseyside, UK; [5]Unilever, Safety and Environmental Assurance Centre, Colworth Science Park, Sharnbrook, UK; [6]JRC Institute for Health and Consumer Protection, Ispra, Italy; [7]University of Antwerp, Zebrafishlab, Wilrijk, Belgium; [8]Agilent Technologies, Washington, DC, USA; [9]The Humane Society of the United States, Washington, DC, USA; [10]Harvard John A. Paulson School of Engineering & Applied Sciences, Cambridge, MA, USA

Abstract

Current efforts in chemical safety are focused on utilizing human *in vitro* or alternatives to animal data in a biological pathway context. However, it remains unclear how biological pathways, and toxicology data developed in that context, can be used to quantitatively facilitate decision-making. The objective of this work is to determine if hypothesis testing using adverse outcome pathways (AOPs) can provide quantitative chemical hazard predictions. Current methods for predicting hazards of chemicals in a biological pathway context were extensively reviewed, specific case studies examined, and computational modeling used to demonstrate quantitative hazard prediction based on an AOP. Since AOPs are chemically agnostic, we propose that AOPs function as hypotheses for how specific chemicals may cause adverse effects via specific pathways. Three broad approaches were identified for testing the hypothesis with AOPs, semi-quantitative weight of evidence, probabilistic, and mechanistic modeling. We then demonstrate how these approaches could be used to test hypotheses using high throughput *in vitro* data and data from alternatives to animal testing. Finally, we discuss standards in development and documentation that would facilitate use in a regulatory context. We conclude that quantitative AOPs provide a flexible hypothesis framework for predicting chemical hazards, which accommodates a wide range of approaches that are useful at many stages and build upon one another to become increasingly quantitative.

1 Introduction

Risk assessors attempt to identify causal relationships between a stressor, for example a chemical, and an outcome of safety and regulatory interest, such as cancer or reproductive impairment, through the integration of a wide range of data and information (NRC, 1983, 2009; Abt et al., 2010). These efforts are facilitated using hypothesis driven approaches and conceptual frameworks to organize or describe data and information related to risk, e.g., sources and pathways of exposure, ADME (adsorption, distribution, metabolism, and elimination), and the health hazard to individuals, communities, or populations (Suter and Cormier, 2011). Several conceptual frameworks have been proposed to support hazard and risk decision-making including mode of action analysis (Sonich-Mullin et al., 2001; USEPA, 2005; Meek et al., 2014), the human relevance framework (Meek et al., 2003), the key events/dose-response framework (Julien et al., 2009), mechanistic effect modeling (Forbes and Calow, 2012), and the adverse outcome pathway (AOP) framework (Ankley et al., 2010). Simultaneously, hazard and risk assessments are evolving to focus on assays relevant to target species (e.g., human rather than rat), *in vitro* test data, 'omics data, and biological pathway perturbations leading to apical level changes (NRC, 2007; Krewski et al., 2014). As a result, many opportunities exist to better support decision-making through development of new approaches for data integration, incorporation of data from emerging technologies, and extrapolating impacts to safety endpoints based on *in vitro* assays and chemical structures.

The AOP framework has been the focus of several research and regulatory organizations (e.g., the U.S. Environmental Protection Agency (EPA), the Organization for Economic Co-operation and Development (OECD), Health Canada, and the European Commission Joint Research Center) as an approach to document and categorize chemical hazards in a biological pathway context. To support this effort, a knowledge base[1] has been created to support the use of AOPs along with guidance for developing AOPs and potential application of AOPs in chemical hazard testing and screening (Meek et al., 2014; Tollefsen et al., 2014; Villeneuve et al., 2014a,b; Becker et al., 2015; Patlewicz et al., 2015; Rovida et al., 2015; Perkins et al., 2015; Groh et al., 2015a,b). However, except for a general discussion on the recent description of the development of a quantitative AOP (qAOP) for aromatase inhibition leading to reproductive dysfunction in fish (Conolly et al., 2017), little guidance exists on how the AOP framework can be used in a quantitative manner to predict impacts on individuals and populations from *in vitro* data or even information derived from clinical studies (OECD, 2013).

The development of AOP-based quantitative models is likely to face similar challenges to those found with biologically-based dose-response models where extensive data and long development times have limited their use in risk assessment (Crump et al., 2010). Indeed, development of a qAOP for aromatase inhibition took several years (Conolly et al., 2017) but this clearly does not have to be the rule. Complicating the prediction of adverse outcomes from AOPs is that the number of different AOPs is likely to be in the hundreds, many of which may interact as networks of pathways either due to the presence of chemical mixtures or activation of multiple events by a single chemical (Garcia-Reyero, 2015). As a result, there is a critical need to examine how quantitative AOPs can be used to meet these challenges, and how their development might be sped up.

Our objective in this paper is to examine how hypothesis testing using quantitative qAOPs can support hazard and risk decisions. We also argue that qAOPs can take many forms, from text-based, descriptive AOPs to probabilistic network models (e.g., Bayesian networks) to mathematical models with high biological fidelity, thereby enabling rapid development of quantitative approaches to examine chemical effects across multiple AOPs.

2 Methods

The state of predicting chemical hazards
Current methods for predicting hazards of chemicals in a biological pathway context were extensively reviewed based on the published literature. Specific case studies were identified and examined for quantitative measures of biological pathway-based hazards.

Mechanistic modeling of a chemical's impact on a population via an AOP
To demonstrate the mechanistic modeling of an AOP, we constructed a coupled component qAOP model representing the AOP for aromatase inhibition leading to reproductive dysfunction using mathematical and probabilistic models representing different KE within the AOP. Briefly, we extended a hypothalamus-pituitary-ovary model of 17β-estradiol synthesis for fathead minnow (*Pimephales promelas*) by Shoemaker et al. (2010) with a liver component that mathematically described the synthesis rate of VTG mRNA and VTG protein in liver controlled by plasma E2 concentrations with a time delay response (Equations S2 and S3[2]). The resulting HPG-liver model was then coupled to the fathead minnow population matrix model of Miller et al. (2007) using a simple linear regression model that describes the relationship between fecundity and plasma VTG concentrations.

3 Results and discussion

3.1 Quantitative AOP frameworks to support decision making
Regulators involved in reviewing chemical use and permissible exposure levels are required to come to a sound and objective scientific judgment as to the potential of that chemical to cause adverse effects on humans and the environment (NRC, 2009). Many regulatory applications use a biological pathway-based hypothesis to examine the causal evidence for human and ecological hazards as part of a risk-based approach (Boobis et al., 2009; Suter and Cormier, 2011; Meek et al., 2014). Since an AOP represents a plausible biological pathway causally leading to an adverse effect, it can be used as a hypothesis that a chemical causes an adverse effect (e.g., cancer, mortality) when used with chemical-specific evidence. An AOP is composed of a molecular initiating event (MIE), where a stressor interacts with a receptor, enzyme or other biomolecule, that in turn causes the activation of a series of measurable events (key events or KE) leading to an adverse outcome (AO) of regulatory interest (Villeneuve et al., 2014a). MIE, KE, and AO are linked by key event relationships (KER) that represent response-response relationships between events that can be used to develop quantitative AOPs.

Arguably the greatest potential use of the AOP framework lies in using it to semi-quantitatively or quantitatively assess, in a transparent manner, the likelihood of a chemical causing an adverse effect, thereby supporting hazard and risk decisions. We anticipate that both existing and future quantitative toxicological models can be incorporated into an AOP framework since AOPs describe biological pathways required for toxicological effects. Three general approaches can be used for hypothesis testing of AOPs in a quantitative manner, ranging from expert judgement-based, requiring limited information, to biological models, which require extensive data and development times (Gust et al., 2015): 1) in a semi-quantitative or quantitative weight of

---

evidence (WOE) manner, where the evidence for a chemical acting through a specific AOP is given a weighted value based on expert opinion and well-documented criteria (e.g., Becker et al., 2017); 2) in a probabilistic manner, where statistical or sufficiency relationships exist between MIE or KE and the AO permits extrapolation from *in vitro* assays or other data to events relevant to safety assessment and of regulatory interest (e.g., Miller et al., 2007; Burgoon et al., 2017); and 3) in a mechanistic manner, where mathematical models of MIE, KE and KER can be used to quantitatively predict the risk of an adverse effect given specified initial conditions (e.g., Conolly et al., 2017). These three areas are complementary and build upon one another to become increasingly more quantitative.

3.2 Semi-quantitative or quantitative weight of evidence qAOPs

A large source of uncertainty can be due to different ways that the same data is valued by different people, which can lead to different assessments of risk (Weed, 2005). This uncertainty between values and conclusions can be bridged through explicit and transparent description of the values and approaches used in assessing data such as clear criteria for how one values toxicity test data. A WOE approach can provide clear criteria and valuations that can be used in reviewing available data supporting hypotheses that a chemical causes an adverse effect through a specific pathway (Weed, 2005). Semi-quantitative and quantitative WOE evaluations incorporate quantitative weighting and numerical assessments of value of data in order to integrate multiple separate lines of evidence into a single value to support decision-making (Linkov et al., 2011; Rhomberg, 2014). The use of transparent WOE approaches can increase certainty in scientific judgement by documenting how data is interpreted and integrated to arrive at a final assessment and overcome reluctance to place values on risk assessments (Linkov and Seager, 2011; Rhomberg, 2014; Linkov, 2015).

A simple and direct semi-quantitative approach for assessing the hypothesis that a chemical causes a health hazard through an AOP uses ranking of confidence in the key event relationships (KER) in the AOP using available evidence. The strength of relationships between events is of particular importance as KER define how a perturbation proceeds from one event to the next (OECD, 2013; Villeneuve et al., 2014a,b). As a result, clearly defined criteria, including biological plausibility, essentiality of each KE, response-response concordance, temporal concordance, incidence concordance (the incidence of upstream KE observations is greater than the downstream KE) and causal evidence can be used to assess the confidence in, or strength of, MIE, KE, KER, indirect KER (the ability to indirectly infer a KE or AO from a non-adjacent event) and ultimately the AOP (OECD, 2013; Becker et al., 2015; Patlewicz et al., 2015). Frameworks for assessing dose response relationships (Simon et al., 2014) or approaches determining points of departure from controls (Thomas et al., 2007; Chepelev et al., 2014; Webster et al., 2015; Burgoon et al., 2017; Farmahin et al., 2016) may also be appropriate for placing a value of strength on the KERs. Once KEs and KERs are scored, either based on strict guidance or ex-

pert judgment, WOE can be used to assess the overall confidence in the ability of a chemical to activate an AOP or its components (Becker et al., 2015; Linkov, 2015). This WOE approach has long been used to evaluate the impact of different factors in environmental risk assessment (Linkov et al., 2006; Linkov and Seager, 2011).

Becker et al. (2017) used this hypothesis driven quantitative WOE approach to assess whether clofibrate induces hepatocarcinomas in rodents by one of two hypothesized AOPs, PPARα activation leading to liver tumors in rodents ((MIE PPARα activation) → (KE of altered cell growth pathways) → (KE of perturbation of cell growth and survival) → (KE of clonal expansion of preneoplastic foci) → (AO of rodent liver tumors)) or Mutagenesis leading to liver tumors in rodents ((MIE DNA reactivity leading to promutagenic adducts/lesions) → (KE of insufficient repair of DNA leading to mutations in key genes) → (KE of perturbation of cell growth and survival) → (KE of clonal expansion of preneoplastic foci) → (AO of rodent liver tumors)). Becker et al. (2017) demonstrated that one could develop a transparent and quantitative assessment of the evidence supporting whether a chemical acted via a specific pathway. An additional value of WOE approaches is the identification of areas where sufficient information exists to develop statistical relationships or where more resources should be invested to better define the pathway.

3.3 Probabilistic quantitative AOPs

Key event relationships and indirect KER describe response-response relationships between events and the outcome and are an essential part of establishing causality in AOPs. Response-response relationships permit development of statistical or probabilistic relationships that enable prediction of the likelihood that a later event or outcome would occur based on changes in an earlier event. The incorporation of statistical or probabilistic relationships into an AOP creates a probabilistic quantitative AOP. This can enhance decision making by requiring less data to support the hypothesis testing for specific chemicals. Probabilistic qAOPs can be composed of predictive relationships that span a few events or an entire AOP and be combined with WOE analyses depending on the application. Probabilistic AOPs can be developed, even when all essential KE have not been identified, if a predictive relationship exists linking events of interest to the AO or an event is causally linked to the AO. Significant statistical linkages between MIE and adverse outcomes have long been used in screening chemicals for the potential to cause hazardous effects. For example, the AOP for membrane disruption (narcosis) leading to respiratory failure is a non-specific toxicity characterized by generalized depression in biological activity that can lead to hypoxia and death for which probabilistic qAOPs exist. Approximately 60% of industrial chemicals are thought to have the potential to exhibit only this mode of action (van Wezel and Opperhuizen, 1995; Volz et al., 2011) given relevant exposures. Non-polar and polar narcotics diffuse into membranes based on lipophilicity, resulting in a significant correlation between a measure of the MIE, hydrophobicity parameter octanol-water partition coefficient or $logK_{ow}$, and the adverse outcome

of respiratory failure (Mackay et al., 2009). Highly predictive quantitative structure activity relationship (QSAR) models have been developed using the predictive relationship of logK$_{ow}$ to non-polar narcotic acute toxicity that are used in hazard screening efforts to test the hypothesis that a chemical acts through the narcosis AOP (Verhaar et al., 1992; Dom et al., 2012; Claeys et al., 2013). QSAR modeling has been widely used to model the effects of chemicals on MIE and has provided substantial support for testing whether or not a specific chemical might initiate an AOP (Allen et al., 2016; Cronin et al., 2017).

The potential of a chemical to cause effects through endocrine signaling is an important hypothesis tested by several different quantitative approaches. Agonism or antagonism of estrogen receptor signaling is involved in several AOPs including reproductive dysfunction in mammals, fish and other species (Ankley et al., 2010; Becker et al., 2015). As there are strong causal linkages between estrogen receptor activation, many efforts have focused upon predicting chemical binding to the estrogen receptor using QSAR (e.g., Tong et al., 1997), machine learning methods (e.g., Zang et al., 2013), molecular docking (e.g., Shao et al., 2004), and other methods. Models of estrogen receptor binding and subsequent activation of estrogen receptor gene expression have also been used to develop a decision model to facilitate hazard identification and prioritization using a combination of structure activity relationships, receptor binding and vitellogenin gene activation assays (Schmieder et al., 2014).

High throughput assay data for estrogen and androgen receptor binding and activation have been used in statistical models

and have been shown to have significant accuracy in predicting *in vivo* endpoints (Rotroff et al., 2013; Cox et al., 2014). Since estrogen receptor (ER) binding and activation is the primary MIE leading to estrogenic adverse outcomes in animals, these models have been proposed as prioritization tools. Judson et al. (2015) tested the hypothesis that a chemical activates the estrogen receptor by extending these efforts to create a partial qAOP describing the MIE of ER binding and activation, the KE of RNA transcription and translation of ER dependent genes, and the KE of ER-dependent cell proliferation. While this does not specifically model biological events, it does incorporate biological assay results that capture these KE by using a mathematical network model that integrates the areas under the curve for assay responses of 18 different high throughput assays to predict potential endocrine agonists and antagonists.

3.4 Probabilistic quantitative AOP networks

Probabilistic approaches such as Bayesian network analysis are well suited to the AOP framework because, like a Bayesian network, an AOP is an intuitive representation of a graphical model that is a formal representation of a joint probability distribution (Koller and Friedman, 2009; Pearl, 2010). Bayesian network approaches have been used to model outcomes in a wide range of complex systems (for review see Weber et al., 2012). Bayesian networks are useful for making probabilistic predictions as to whether one or multiple hypotheses are likely to be true, provide diagnostic analysis of evidence available for decisions, and the ability to update calculations based on new evidence such as ad-
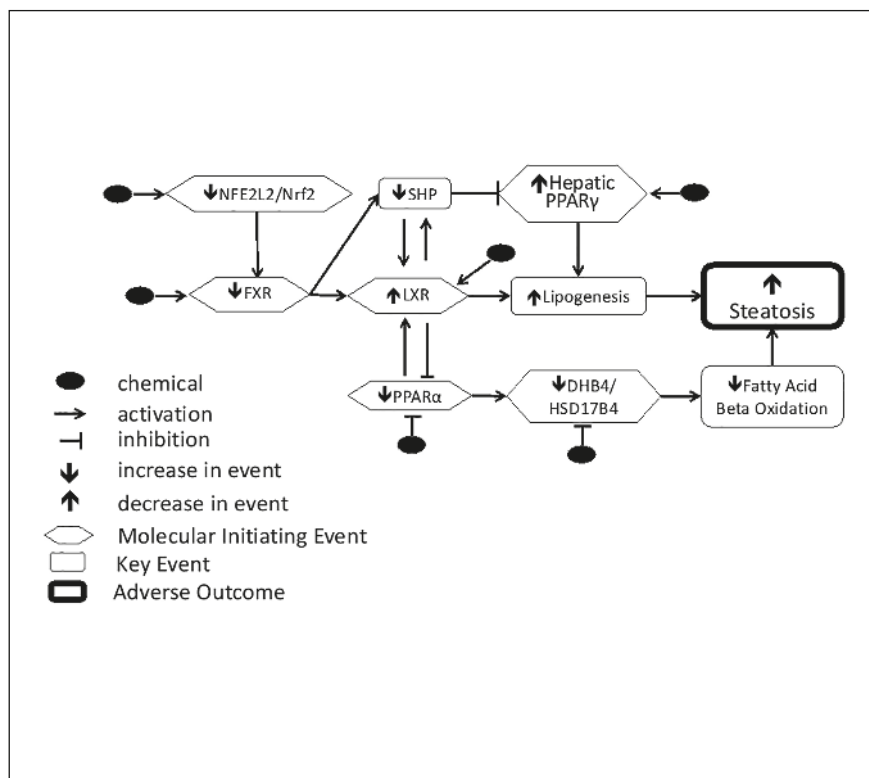
Fig. 1: Conceptual model of a quantitative steatosis adverse outcome network

Six possible AOPs are initiated by molecular initiating events (MIE, hexagon boxes) leading to the adverse outcome steatosis. The probability that one event leads to another is represented by an arrow. The probability an event will interfere with or inhibit another event is represented by a line and bar. The probability that any one AOP will result in steatosis is represented by the joint probability distribution across that AOP. Possible crosstalk between different AOPs is revealed in the network. The effect of complex mixtures could be assessed by examining the joint probability distribution across the entire network given the available data. DHB4/HSD17B4, hydroxysteroid (17-β) dehydrogenase 4; FXR, farnasoid X receptor; LXR, liver X receptor; NFE2L2/Nrf2, nuclear factor, erythroid 2 like 2; PPAR, peroxisome proliferator-activator receptor; SHP, small heterodimer partner

ditional *in vitro* tests. For example, based on empirical evidence showing that various events in the AOP are predictive of the potential of a chemical to be a skin sensitizer, several non-animal test methods that measure the impact of chemical sensitizers on these key events have been developed (Liebsch et al., 2011; Maxwell et al., 2014). However, individually, the assays are inconsistent in predicting the relative potency of a chemical. Consequently, integrated testing strategies for the AOP for skin sensitization caused by covalent binding to proteins have been developed where Bayesian network models incorporate *in vitro* assays representing MIE and KE events to predict the potency of a chemical in inducing a response in a local lymph node assay (Pirone et al., 2014). The value of the skin sensitization Bayesian network is that it provides a probabilistic estimation of the sensitization potential of a chemical and allows a quantitative examination of whether addition of more tests would improve the predictive ability of the framework. A further example is the toxicokinetic/toxicodynamic modeling approach taken by MacKay et al. (2013), who have extended an existing skin bioavailability model (Davies et al., 2011) to estimate the probability of allergy in a given human population to be predicted using a toxicodynamic model of skin protein haptenation, DC antigen presentation, and CD8[+] T cell activation for application in skin sensitization risk assessment.

A major limitation of linear AOPs is that real world risks are due to multiple factors such as diseases caused by interactions between susceptible genotypes and the presence of certain metals, cross-talk between pathways, chemicals that initiate multiple AOPs, or complex mixtures where an organism may be exposed to multiple chemicals. The flexibility of Bayesian networks is also valuable in examining multiple interacting variables and networks of interacting AOPs. For example, non-alcoholic liver steatosis can be caused via several AOPs creating a complex network where multiple interactions could contribute to steatosis (Fig. 1; Hashimoto et al., 2000; Reddy, 2001; Grefhorst et al., 2002; Pineda Torra et al., 2003; Kay et al., 2011; Sharif et al., 2014). Understanding the potential contribution of multiple AOPs in the presence of chemical mixtures can be very challenging. Given the appropriate experimental data and formal relationship criteria (OECD, 2013; Becker et al., 2015), statistical relationships could be developed to enable construction of a probabilistic network that captures potential crosstalk between pathways. For example, the AOP network for steatosis suggests that a mixture of PPARγ agonists and FXR antagonists could lead to a greater risk of steatosis by increasing lipogenesis through removing feedback repression of LXRα expression (Goodwin et al., 2000; Lu et al., 2000) and increasing PPARγ dependent expression of enzymes involved in lipogenesis (Morán-Salvador et al., 2011). If placed in a Bayesian network context, multiple hypotheses can be tested.

3.5 Mechanistic quantitative AOPs

While probabilistic qAOPs are clearly useful in estimating if an adverse outcome may occur given available data, they generally do not explicitly incorporate mechanisms of action and fail to account for regulatory and feedback control mechanisms

that underlay compensatory responses. Feedback regulation is a central component of many biological systems (Bhalla and Iyengar, 1999; Avraham and Yarden, 2011; Cowan et al., 2014) including the hypothalamus-pituitary-gonadal (HPG) axis (Norris and Carr, 2013) and the hypothalamus-pituitary-thyroid axis (Chiamolera and Wondisford, 2009; Carr and Patiño, 2011). Mechanistic qAOPs are typically more complex and time-consuming to construct than WOE or probabilistic qAOPs, include (first order) mathematical relationships, and generally represent a more accurate biological model (e.g., Conolly et al., 2017). Mechanistic modeling also facilitates a quantitative estimate of uncertainty in the risk assessment. Mechanistic qAOPs can also include toxicokinetic and toxicodynamic modeling that explicitly capture the details on absorption, distribution, metabolism, and elimination of chemicals. This level of modeling requires large amounts of data and remains a long-term effort for most current AOPs.

3.6 Hypothesis testing with mechanistic quantitative AOPs

The AOP for inhibition of aromatase leading to reproductive dysfunction in fish (Becker et al., 2015) provides a well characterized pathway with which to highlight essential features of a mechanistic qAOP (Conolly et al., 2017). Here, we constructed a mechanistic qAOP for aromatase inhibition using mathematical and probabilistic models representing different KE within the aromatase inhibition AOP (Fig. 2, details of model development and predictions in supplementary file[2]). The biological effects of aromatase inhibition in fathead minnow have been extensively studied and sufficient data exists to support mechanistic model development (Ankley et al., 2002; Villeneuve et al., 2009, 2013). Fadrozole (FAD) is a model endocrine disruptor that specifically inhibits aromatase, an enzyme that catalyzes the final step in the synthesis of estradiol (E2), an estrogen required for reproductive function (Browne et al., 1991). Breen et al. (2007) developed a metabolic model of *ex vivo* fathead minnow ovary slice assays converting cholesterol to testosterone and estradiol that described the MIE of inhibition of aromatase by the inhibitor fadrozole and KE1 – a decrease in estradiol synthesis. Shoemaker et al. (2010) extended this model to incorporate endocrine signaling feedback control from the ovary to the hypothalamus/pituitary, the luteinizing hormone / luteinizing hormone receptor signaling cascade, regulation of the steroidogenic acute response protein responsible for cholesterol transport into mitochondria, and critical transcription factors into a dynamic mathematical model to predict the effects of fadrozole exposure on plasma E2 concentrations (Fig. 2). The Shoemaker model was able to accurately predict effects of aromatase inhibition on testosterone and E2 production by fathead minnow in the presence of 50 μg/l FAD over 6, 12, and 24 h (Shoemaker et al., 2010). The model is also generally predictive of plasma E2 concentration behavior after 8 days exposure to FAD, although it fails to predict compensation for FAD inhibition of aromatase at low (3 μg/l) concentrations and slow recovery of normal E2 levels after removal of FAD (Fig. 3a). This is consistent with the findings of Breen et al. (2013)
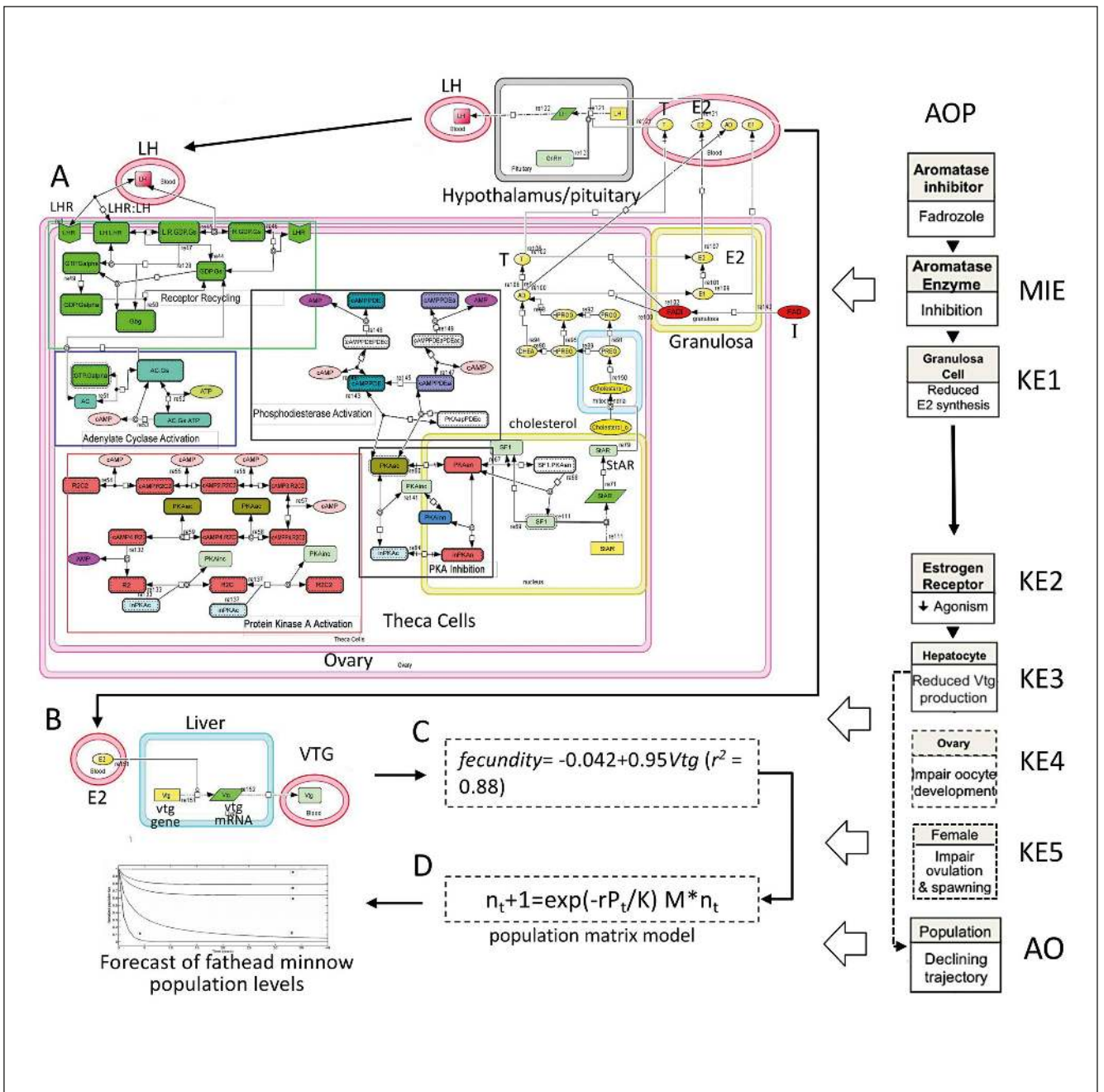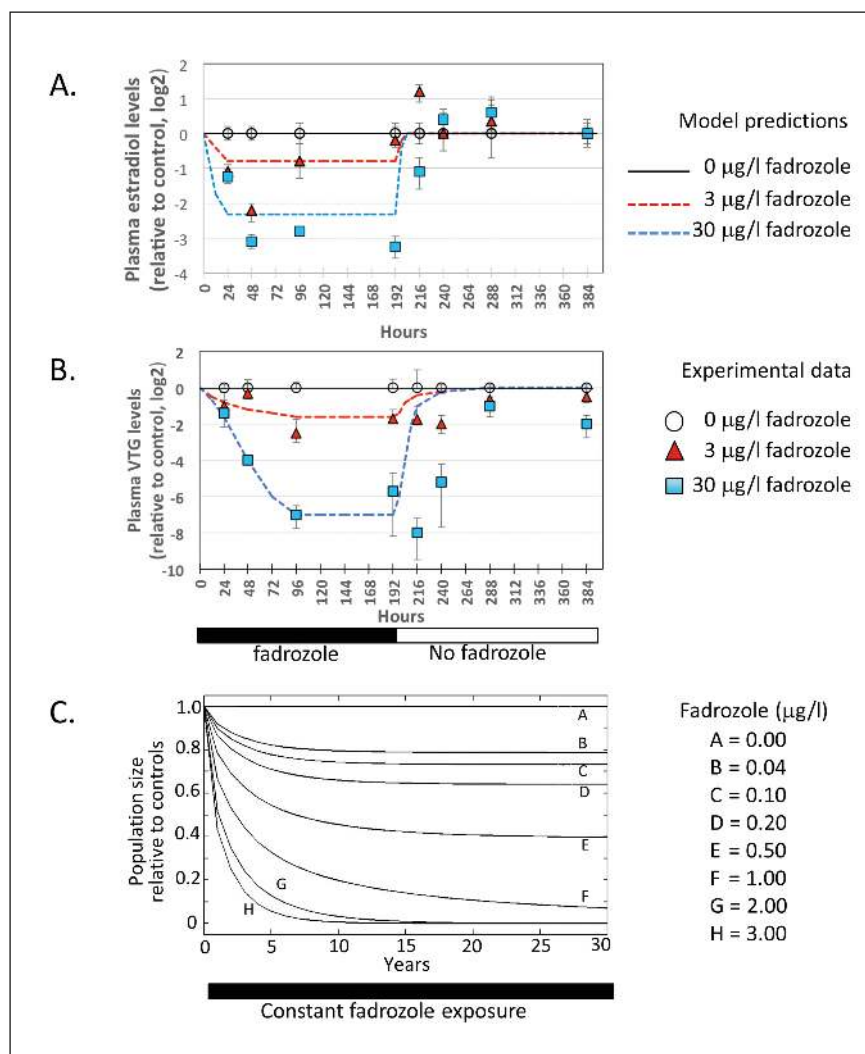
Fig. 2: Linkage of multiple models to create a mechanistic qAOP model for aromatase inhibition leading to reproductive dysfunction

The AOP for aromatase inhibition leading to reproductive dysfunction is diagrammed on the right of the figure with arrowhead indicating equivalent sections of the mechanistic qAOP. (A) A mechanistic hypothalamus-pituitary-gonad (HPG) model that simulates steroidogenesis (yellow) and conversion of testosterone (T) into estradiol (E2) by aromatase, a negative feedback to the hypothalamus/pituitary where decreasing levels of E2 cause increased synthesis of luteinizing hormone (LH; pink), LH binds to the LH receptor in the ovary and stimulates a cAMP cascade (green, purple) resulting in phosphorylation (light green, red) of transcription factor SF1, increased transport of cholesterol into mitochondria by steroid acute response protein and ultimately increased synthesis of E2. I represents a chemical aromatase inhibitor, here fadrozole. (B) Vitellogenin liver compartment model. (C) Model relating vitellogenin levels to fecundity. (D) Density-dependent population matrix model.

that additional regulatory or biological elements not explicitly described in the AOP or the model exist that influence fathead minnow responses to FAD.

We extended the Shoemaker HPG model to create a HPG-liver model that includes KE2, a decrease in estrogen receptor agonism, and KE3, reduced VTG production in liver, by mathematically describing the synthesis rate of VTG mRNA controlled by plasma E2 concentrations with a time delay response (Equation S2[2]). The process of translation of VTG protein was also described with a time delay function in a separate equation (Equation S3[2]). The resulting HPG-liver model accurately predicted plasma VTG levels in the presence of low (3 µg/l) and high (30 µg/l) concentrations of FAD (Fig. 3b). As with predictions of plasma E2 levels, predicted plasma VTG levels rapidly returned to normal, whereas observed plasma VTG levels took significantly longer to return to normal, indicating that additional biological mechanisms need to be incorporated into the models to accurately simulate the impact of FAD on plasma VTG levels.

Since a decrease in plasma VTG levels in female fathead minnows is highly correlated to fecundity, KE5, impaired ovulation and spawning (Fig. 2), can be modelled using a simple linear regression model developed by Miller et al. (2007) based on 21-day reproductive studies with different chemical stressors that describes the relationship between fecundity and plasma VTG concentrations relative to untreated females (Equation S4[2]). The population matrix model of Miller et al. (2007) can then be used to create a complete model of the AOP from aromatase inhibition through the population level outcome by linking the fecundity model, KE5, to population level effects, the AO (see supplementary file[2]). We used the complete model (Fig. 2) to predict population trajectories for various levels of constant FAD exposure (Fig. 3). As observed by Miller et al. (2007), depression of VTG levels results in a decline of population trajectories, relative to no FAD exposure, which stabilize at lower population levels. Concentrations of FAD above 1µg/l produced catastrophic effects resulting in the total collapse of the population (Fig. 3).

Fig. 3: Quantitative AOP modeling of KE and AO responses for aromatase inhibition leading to reproductive dysfunction in fish in the presence of fadrozole

Calculated responses of (A) plasma estradiol levels and (B) plasma vitellogenin levels in fathead minnow exposed for 8 days (192 h) to fadrozole followed by 8 days of recovery; black line, control; blue line, 30 µg/l fadrozole; red line, 30 µg/l fadrozole. Symbols indicate experimental data points (adapted from Villeneuve et al., 2010); circle, control; red triangle, 3 µg/l fadrozole; blue box, 30 µg/l fadrozole. Y axis indicates the normalized value relative to control on a log2 scale. The bar below graphs A and B depicts exposure to fadrozole (black bar) and recovery without fadrozole (white bar). (C) Relative population trajectory forecasted for fathead minnow (*Pimephales promelas*) exposed to different levels of fadrazole. Relative population becomes stable at lower capacity (78%) upon exposure to 0.04 µg/l fadrozole while capacity reduces to 40% at exposure to 0.5 µg/l fadrozole. Population becomes zero after 15 and 10 years under exposure to 2 and 3 µg/l fadrozole, respectively. The black bar beneath graph C indicates constant exposure to fadrozole.

The qAOP model described here did not model compensatory behavior found when fathead minnow were exposed to low doses of fadrozole (Villeneuve et al., 2009, 2013). Therefore, while the mechanistic qAOP incorporated many of the biological elements essential to describing the aromatase inhibition AOP, additional regulatory mechanisms, such as those proposed by Breen et al. (2013), are still required to capture dynamic behaviors at low doses of chemical inhibitors.

3.7 Data standards for development of qAOPs

An essential component of all three approaches to qAOP development is the need for extensive review and documentation of available data and literature. While expert judgment will remain central to AOP development, including qAOPs, transparency in how relationships and values were derived in addition to the reproducibility of WOE, probabilistic, and mechanistic values, will be critical to gain acceptance for an AOP/qAOP. Systematic review has emerged as one approach to help ensure transparency and begin to provide guidelines for reproducibility. We view the systematic review standards proposed by the Institute of Medicine (IOM, 2011) to be transferable and helpful in development of AOPs, especially in development and documentation of WOE, probabilistic, and mechanistic qAOPs.

Currently, AOPs are documented and reviewed via the AOP-Wiki[3] under the auspices of the Organization for Economic Cooperation and Development (OECD), which also implements an internal and external expert review of AOPs[4]. Many of the elements of the systematic review standards can be easily incorporated into the AOP-Wiki to provide transparency in development of AOPs for risk assessment. Here we briefly outline some of our vision of what standards for AOPs might look like, along with some reasoning for their importance.

Standard 1: Initiating an AOP systematic review project
The key element of this standard is to establish the systematic review protocol. This includes describing exactly what the adverse outcome is that will be assessed, in what species the literature search will be conducted, describing the literature search and data extraction strategies, identifying any conflicts of interest that team members have, and describing the study quality protocols. The purpose of this standard is to ensure a rigorous protocol is in place, and that the investigators have given this some thought. Ideally, the project team would put their protocol out for comment for a short period of time and incorporate the feedback they obtained prior to starting the project.

Standard 2: Execute the protocol
The AOP team executes the searches, performs study quality analyses (this might include a risk of bias analysis), and extracts the data and places it into a centralized database. Any deviations from the protocol must be noted. To ensure transparency, the team will need to document reasons for including and excluding studies. The goal is for teams to provide sufficient documentation of their reasoning so that another scientist trained in the field can understand, not necessarily agree, how the team reached the conclusions. This is where the use of standardized questionnaires or score sheets becomes valuable and tools such as the National Institutes of Health's Office of Health Assessment and Translation (OHAT) Risk of Bias Tool[5] or the Systematic Omics Analysis Review (SOAR) tool (McConnell et al., 2014) are helpful.

Standard 3: AOP synthesis
Here, the information from individual studies is synthesized into an overall AOP on the AOP-Wiki. Evidence for events and relationships should be assessed via weight of evidence frameworks as described in Becker et al. (2015). We strongly advocate the use of meta-analysis and statistical methods to integrate evidence from multiple studies. Publication bias across the databases can be identified using a trim-and-fill plot (Duval and Tweedie, 2000). Analyses using Galbraith plots can help identify if there are differences in standard errors from different studies as a function of effect size (Galbraith, 1990). In an ideal situation, where multiple studies all report statistical analyses for the same potential key event, meta-analyses could be used to ascertain the strength of the evidence that a potential key event is or is not involved. For instance, where studies have used gene knock-out models, several individual studies might report that the knock-out of a protein representing a key event is necessary to block the adverse outcome, while a few studies may contradict this view. If all the studies had reasonable approaches, then the data could be integrated using Bayesian analysis, thereby permitting an objective determination of whether the data supports the protein being used as a key event or not. The analysis and the results would then become part of the argument for/against inclusion of the key event.

Standard 4: Submit the AOP
For AOPs/qAOPs to be useful, they need to be made publicly available so that others can use them. The AOP-Wiki is an excellent resource for making AOPs available and, potentially, qAOPs in the future. If an AOP/qAOP is made part of the OECD work plan, it will be peer-reviewed by experts. Although we do not see requiring AOPs/qAOPs to be made publicly available as part of the standard, we do see the need for data standards to share them. On the computational side, the AOP ontology[6] allows AOPs (and qAOPs soon) to be exchanged between computers, and to facilitate computational analyses. We would also like to have methods that can translate future exchange formats with other existing standards for systems biology, such as systems biology mark up language.

Although these systematic review standards are currently a work in progress, we feel they will help improve the quality of

---

[3] https://aopwiki.org

[4] http://www.oecd.org/chemicalsafety/testing/adverse-outcome-pathways-molecular-screening-and-toxicogenomics.htm

[5] https://ntp.niehs.nih.gov/ntp/ohat/pubs/riskofbiastool_508.pdf

[6] https://github.com/DataSciBurgoon/aop-ontology

future AOPs/qAOPs, and will facilitate meaningful dialogue, discussion, and the identification of best practices as we continue to develop these tools into the future.

4 Conclusion

These examples of qAOPs demonstrate several important points. 1) Quantitative AOPs can be developed at many different stages (weight of evidence to mechanistic) depending on the amount of information available and the intended use. 2) qAOPs can be based on a combination of multiple modeling approaches including deterministic mathematical models and statistical models. 3) Regulatory mechanisms are not explicitly included as a separate element in an AOP, yet are essential for developing accurate mechanistic qAOPs. If AOPs are to be useful in a mechanistic context, then the relationships between KE must be understood and examined in the context of regulatory interactions important in moving from one KE to another. 4) The application of qAOPs in a full risk assessment requires consideration of chemical exposure since the amount of chemical present to initiate an AOP is dependent upon exposure conditions, chemical bioavailability, adsorption, distribution, metabolism, or excretion parameters. Therefore, an AOP coupled to exposure events provides an intuitive framework for assessing the probability that exposure to a chemical will cause interaction with a MIE, thereby affecting molecular processes at the cellular level leading to adverse outcomes such as cancer, decreased survival and reproduction, or declines in population size and growth. Future use in risk assessment and decision-making will require a level of confidence in model predictions at least as high as the currently accepted methods. Further efforts will be required to benchmark the output from these mathematical models against relevant datasets, such as human clinical datasets (e.g., diagnostic patch test data) or ecological species where appropriate (Maxwell et al., 2014), although the availability of such data may be limited.

We anticipate that in the near term, as high confidence AOPs begin to emerge, we will see the rise of computational predictions based on AOPs for limited decision-making (e.g., prioritization, screening, and hazard identification), where assay data will be applied to these high confidence AOPs to make hazard predictions. Ultimately, this will lay the foundation for more quantitative use of AOPs in support of predictive toxicology and preliminary risk estimates that can be used by risk assessors as the starting point for future chemical risk assessments.

References

Abt, E., Rodricks, J. V., Levy, J. I. et al. (2010). Science and decisions, advancing risk assessment. *Risk Anal 30*, 1028-1036. doi:10.1111/j.1539-6924.2010.01426.x

Allen, T. E., Liggi, S., Goodman, J. M. et al. (2016). Using molecular initiating events to generate 2D structure-activity relationships for toxicity screening. *Chem Res Toxicol 29*, 1611-1627. doi:10.1021/acs.chemrestox.6b00101

Ankley, G. T., Kahl, M. D., Jensen, K. M. et al. (2002). Evaluation of the aromatase inhibitor fadrozole in a short-term reproduction assay with the fathead minnow (Pimephales promelas). *Toxicol Sci 67*, 121-130. doi:10.1093/toxsci/67.1.121

Ankley, G.T., Bennett, R. S., Erickson, R. J. et al. (2010). Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. *Environ Toxicol Chem 29*, 730-741. doi:10.1002/etc.34

Avraham, R. and Yarden, Y. (2011). Feedback regulation of EGFR signaling, decision making by early and delayed loops. *Nat Rev Mol Cell Bio 12*, 104-117. doi:10.1038/nrm3048

Becker, R. A., Ankley, G. T., Edwards, S. W. et al. (2015). Increasing scientific confidence in adverse outcome pathways: Application of tailored Bradford-Hill considerations for evaluating weight of evidence. *Regul Toxicol Pharmacol 72*, 514-537. doi:10.1016/j.yrtph.2015.04.004

Becker, R. A., Dellarco, V. Seed, J. et al. (2017). Quantitative weight of evidence to assess confidence in potential modes of action. *Regul Toxicol Pharmacol 86*, 205-220. doi:10.1016/j.yrtph.2017.02.017

Bhalla, U. S. and Iyengar, R. (1999). Emergent properties of networks of biological signaling pathways. *Science 283*, 381-387. doi:10.1126/science.283.5400.381

Boobis, A. R., Daston, G. P. Preston, R. J. and Olin, S. S. (2009). Application of key events analysis to chemical carcinogens and noncarcinogens. *Crit Rev Food Sci Nutr 49*, 690-707. doi:10.1080/10408390903098673

Breen, M., Villeneuve, D., Breen, M. et al. (2007). Mechanistic computational model of ovarian steroidogenesis to predict biochemical responses to endocrine active compounds. *Ann Biomed Eng 35*, 970-981. doi:10.1007/s10439-007-9309-7

Breen, M., Villeneuve, D. L., Ankley, G. T. et al. (2013). Developing predictive approaches to characterize adaptive responses of the reproductive endocrine axis to aromatase inhibition: II. Computational modeling. *Toxicol Sci 133*, 234-247. doi:10.1093/toxsci/kft067

Browne, L. J., Gude, C., Rodriguez, H. et al. (1991). Fadrozole hydrochloride: A potent, selective, nonsteroidal inhibitor of aromatase for the treatment of estrogen-dependent disease. *J Med Chem 34*, 725-736. doi:10.1021/jm00106a038

Burgoon, L. D., Druwe, I. L., Painter, K. and Yost., E. E. (2017). Using in vitro high-throughput screening data for predicting benzo[k]fluoranthene human health hazards. *Risk Anal 37*, 280-290. doi:10.1111/risa.12613

Carr, J. A. and Patiño, R. (2011). The hypothalamus-pituitary-thyroid axis in teleosts and amphibians, endocrine disruption and its consequences to natural populations. *Gen Comp Endocrinol 170*, 299-312. doi:10.1016/j.ygcen.2010.06.001

Chepelev, N. L., Meek, M. E. and Yauk, C. L. (2014). Application of benchmark dose modeling to protein expression data in the development and analysis of mode of action/adverse outcome pathways for testicular toxicity. *J Appl Toxicol 34*, 1115-1121. doi:10.1002/jat.3071

Chiamolera, M. I. and Wondisford, F. E. (2009). Thyrotropin-releasing hormone and the thyroid hormone feedback mechanism. *Endocrinology 150*, 1091-1096. doi:10.1210/en.2008-1795

Claeys, L., Iaccino, F., Janssen, C. R. et al. (2013). Development and validation of a quantitative structure-activity relationship for chronic narcosis to fish. *Environ Toxicol Chem 32*, 2217-2225. doi:10.1002/etc.2301

Conolly, R. B., Ankley, G. T., Cheng, W. et al. (2017). Quantitative adverse outcome pathways and their application to predictive toxicology. *Environ Sci Technol 51*, 4661-4672. doi:10.1021/acs.est.6b06230

Cowan, N. J., Ankarali, M. M., Dyhr, J. P. et al. (2014). Feedback control as a framework for understanding tradeoffs in biology. *Integr Comp Biol 54*, 223-237. doi:10.1093/icb/icu050

Cox, L. A. T., Popken, D., Marty, M. S. et al. (2014). Developing scientific confidence in HTS-derived prediction models: Lessons learned from an endocrine case study. *Regul Toxicol Pharmacol 69*, 443-450. doi:10.1016/j.yrtph.2014.05.010

Cronin, M. T. D., Enoch, S. J., Mellor, C. L. (2017). In silico prediction of organ level toxicity: Linking chemistry to adverse effects. *Toxicol Res 33*, 173-182. doi:10.5487/TR.2017.33.3.173

Crump, K. S., Chen, C., Chiu, W. A. et al. (2010). What role for biologically based dose-response models in estimating low-dose risk? *Environ Health Perspect 118*, 585-588. doi:10.1289/ehp.0901249

Davies, M., Pendlington, R. U., Roper, C. S. et al. (2011). Determining epidermal dispositions kinetics for use in an integrated nonanimal approach to skin sensitization risk assessment. *Toxicol Sci 119*, 308-318. doi:10.1093/toxsci/kfq326

Dom, N., Vergauwen, L., Vandenbrouck, T. et al. (2012). Physiological and molecular effect assessment versus physico-chemistry based mode of action schemes: Daphnia magna exposed to narcotics and polar narcotics. *Environ Sci Technol 46*, 10-18. doi:10.1021/es201095r

Duval, S. and Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics 56*, 455-463. doi:10.1111/j.0006-341X.2000.00455.x

Farmahin, R., Williams, A., Kuo, B. et al. (2016). Recommended approaches in the application of toxicogenomics to derive points of departure for chemical risk assessment. *Arch Toxicol 91*, 2045-2065. doi:10.1007/s00204-016-1886-5

Forbes, V. E. and Calow, P. (2012). Promises and problems for the new paradigm for risk assessment and an alternative approach involving predictive systems models. *Environ Toxicol Chem 31*, 2663-2671. doi:10.1002/etc.2009

Galbraith, R. F. (1990). The radial plot: Graphical assessment of spread in ages. *Int J Rad Appl Instrum D Nuc Tracks Rad Meas 17*, 207-214. doi:10.1016/1359-0189(90)90036-W

Garcia-Reyero, N. (2015). Are adverse outcome pathways here to stay? *Environ Sci Technol 49*, 3-9. doi:10.1021/es504976d

Goodwin, B., Jones, S. A., Price, R. R. et al. (2000). A regulatory cascade of the nuclear receptors FXR, SHP-1, and LRH-1 represses bile acid biosynthesis. *Mol Cell 6*, 517-526. doi:10.1016/S1097-2765(00)00051-4

Grefhorst, A., Elzinga, B. M., Voshol, P. J. et al. (2002). Stimulation of lipogenesis by pharmacological activation of the liver X receptor leads to production of large, triglyceride-rich very low density lipoprotein particles. *J Biol Chem 277*, 34182-34190. doi:10.1074/jbc.M204887200

Groh, K. J., Carvalho, R. N., Chipman, J. K. et al. (2015a). Development and application of the adverse outcome pathway framework for understanding and predicting chronic toxicity: I. Challenges and research needs in ecotoxicology. *Chemosphere 120*, 764-777. doi:10.1016/j.chemosphere.2014.09.068

Groh, K. J., Carvalho, R. N., Chipman, J. K. et al. (2015b). Development and application of the adverse outcome pathway framework for understanding and predicting chronic toxicity: II. A focus on growth impairment in fish. *Chemosphere 120*, 778-792. doi:10.1016/j.chemosphere.2014.10.006

Gust, K. A., Collier, M. D., Mayo, M. et al. (2015). Limitations of toxicity characterization in life cycle assessment – Can adverse outcome pathways provide a new foundation? *Integr Environ Assess Manag 12*, 580-590. doi:10.1002/ieam.1708

Hashimoto, T., Cook, W. S., Qi, C. et al. (2000). Defect in peroxisome proliferator-activated receptor alpha-inducible fatty acid oxidation determines the severity of hepatic steatosis in response to fasting. *J Biol Chem 275*, 28918-28928. doi:10.1074/jbc.M910350199

IOM – Institute of Medicine (2011). *Finding What Works in Health Care: Standards for Systematic Reviews*. Washington, DC, USA: National Academies Press.

Julien, E., Boobis, A. R., Olin, S. S. and ILSI Research Foundation Threshold Working Group (2009). The key events dose-response framework: A cross-disciplinary mode-of-action based approach to examining dose-response and thresholds. *Crit Rev Food Sci Nutr 49*, 682-689. doi:10.1080/10408390903110692

Judson, R. S., Magpantay, F. M., Chickarmane, V. et al. (2015). Integrated model of chemical perturbations of a biological pathway using 18 in vitro high-throughput screening assays for the estrogen receptor. *Toxicol Sci 148*, 137-154. doi:10.1093/toxsci/kfv168

Kay, H. Y., Kim, W. D., Hwang, S. J. et al. (2011). Nrf2 inhibits LXRα-dependent hepatic lipogenesis by competing with FXR for acetylase binding. *Antioxid Redox Signal 15*, 2135-2146. doi:10.1089/ars.2010.3834

Koller, D. and Friedman, N. (2009). Probabilistic Graphical Models: Principles and Techniques. MIT Press.

Krewski, D., Westphal, M., Andersen, M. E. et al. (2014). A framework for the next generation of risk science. *Environ Health Perspect 122*, 796-805. doi:10.1289/ehp.1307260

Liebsch, M., Grune, B., Seiler, A. et al. (2011). Alternatives to animal testing, current status and future perspectives. *Arch Toxicol 85*, 841-858. doi:10.1007/s00204-011-0718-x

Linkov, I., Satterstrom, F. K., Kiker, G. et al. (2006). Multicriteria decision analysis: A comprehensive decision approach for management of contaminated sediments. *Risk Anal 26*, 61-78. doi:10.1111/j.1539-6924.2006.00713.x

Linkov, I. and Seager, T. P. (2011). Coupling multi-criteria decision analysis, life-cycle assessment, and risk assessment for emerging threats. *Environ Sci Technol 45*, 5068-5074. doi:10.1021/es100959q

Linkov, I., Welle, P., Loney, D. et al. (2011). Use of multicriteria decision analysis to support weight of evidence evaluation. *Risk Anal 31*, 1211-1225. doi:10.1111/j.1539-6924.2011.01585.x

Linkov, I. (2015). From "weight of evidence" to quantitative data integration using multicriteria decision analysis and Bayesian methods. *ALTEX 32*, 3-8. doi:10.14573/altex.1412231

Lu, T. T., Makishima, M., Repa, J. J. et al. (2000). Molecular basis for feedback regulation of bile acid synthesis by nuclear receptors. *Mol Cell 6*, 507-515. doi:10.1016/S1097-2765(00)00050-2

Mackay, D., Arnot, J. A., Petkova, E. P. et al. (2009). The physicochemical basis of QSARs for baseline toxicity. *SAR QSAR Environ Res 20*, 393-414. doi:10.1080/10629360902949153

MacKay, C., Davies, M., Summerfield, V. and Maxwell, G. (2013). From pathways to people: Applying the adverse outcome pathway (AOP) for skin sensitization to risk assessment. *ALTEX 30*, 473-486. doi:10.14573/altex.2013.4.473

Maxwell, G., MacKay, C., Cubberley, R. et al. (2014). Applying the skin sensitisation adverse outcome pathway (AOP) to quantitative risk assessment. *Toxicol In Vitro 28*, 8-12. doi:10.1016/j.tiv.2013.10.013

McConnell, E. R., Bell, S. M., Cote, I. et al. (2014). Systematic omics analysis review (SOAR) tool to support risk assessment. *PLoS One 9*, e110379. doi:10.1371/journal.pone.0110379

Meek, M. E., Bucher, J. R., Cohen, S. M. et al. (2003). A framework for human relevance analysis of information on carcinogenic modes of action. *Crit Rev Toxicol 33*, 591-653. doi:10.1080/713608373

Meek, M. E., Boobis, A., Cote, I. et al. (2014). New developments in the evolution and application of the WHO/IPCS framework on mode of action/species concordance analysis. *J Appl Toxicol 34*, 1-18. doi:10.1002/jat.2949

Miller, D. H., Jensen, K. M., Villeneuve, D. L. et al. (2007). Linkage of biochemical responses to population-level effects: A case study with vitellogenin in the fathead minnow (Pimephales promelas). *Environ Toxicol Chem 26*, 521-527. doi:10.1897/06-318R.1

Morán-Salvador, E., López-Parra, M., García-Alonso, V. et al. (2011). Role for PPARγ in obesity-induced hepatic steatosis as determined by hepatocyte- and macrophage-specific conditional knockouts. *FASEB J 25*, 2538-2550. doi:10.1096/fj.10-173716

Norris, D. O. and Carr, J. A. (2013). *Vertebrate Endocrinology*. 5th edition. Academic Press. ISBN 9780123964656.

NRC – National Research Council (1983). *Risk Assessment in the Federal Government: Managing the Process*. Washington, DC: The National Academies Press. doi:10.17226/366

NRC (2007). *Toxicity Testing in the 21st Century: A Vision and a Strategy*. Washington, DC: The National Academies Press. doi:10.17226/11970

NRC (2009). *Science and Decisions: Advancing Risk Assessment*. Washington, DC: The National Academies Press. doi:10.17226/12209

OECD (2013). Revised Guidance Document for Developing and Assessing Adverse Outcome Pathways (AOPs). 2nd edition. ENV/JM/MONO(2013)6. *Series on Testing and Assessment No. 184*. OECD Publishing, Paris. https://bit.ly/1RJXOEx

Patlewicz, G., Simon, T. W., Rowlands, J. C. et al. (2015). Proposing a scientific confidence framework to help support the application of adverse outcome pathways for regulatory purposes. *Regul Toxicol Pharmacol 71*, 463-477. doi:10.1016/j.yrtph.2015.02.011

Pearl, J. (2010). An introduction to causal inference. *Int J Biostat 6*, Article 7. doi:10.2202/1557-4679.1203

Perkins, E. J., Antczak, P. Burgoon, L. et al. (2015). Adverse outcome pathways for regulatory applications, examination of four case studies with different degrees of completeness and scientific confidence. *Toxicol Sci 148*, 14-25. doi:10.1093/toxsci/kfv181

Pineda Torra, I., Claudel, T., Duval, C. et al. (2003). Bile acids induce the expression of the human peroxisome proliferator-activated receptor α gene via activation of the farnesoid X receptor. *Mol Endocrinol 17*, 259-272. doi:10.1210/me.2002-0120

Pirone, J. R., Smith, M., Kleinstreuer, N. C. et al. (2014). Open source software implementation of an integrated testing strategy for skin sensitization potency based on a Bayesian network. *ALTEX 31*, 336-340. doi:10.14573/altex.1310151

Reddy, J. K. (2001). Nonalcoholic steatosis and steatohepatitis. III. Peroxisomal beta-oxidation, PPAR alpha, and steatohepatitis. *Am J Physiol Gastrointest Liver Physiol 281*, G1333-9. doi:10.1152/ajpgi.2001.281.6.G1333

Rhomberg, L. (2014). Hypothesis-based weight of evidence: An approach to assessing causation and its application to regulatory toxicology. *Risk Anal 35*, 1114-1124. doi:10.1111/risa.12206

Rotroff, D. M., Dix, D. J., Houck, K. A. et al. (2013). Using in vitro high throughput screening assays to identify potential endocrine-disrupting chemicals. *Environ Health Perspect 121*, 7-14. doi:10.1289/ehp.1205065

Rovida, C., Asakura, S., Daneshian, M. et al. (2015). Toxicity testing in the 21st century beyond environmental chemicals. *ALTEX 32*, 171-181. doi:10.14573/altex.1506201

Schmieder, P. K., Kolanczyk, R. C., Hornung, M. W. et al. (2014). A rule-based expert system for chemical prioritization using effects-based chemical categories. *SAR QSAR Environ Res 25*, 253-287. doi:10.1080/1062936X.2014.898691

Shao, D., Berrodin, T. J., Manas, E. et al. (2004). Identification of novel estrogen receptor α antagonists. *J Steroid Biochem Mol Biol 88*, 351-360. doi:10.1016/j.jsbmb.2004.01.007

Sharif, A. l., Alov, M. P., Vitcheva, V. et al. (2014). Modes-of-action related to repeated dose toxicity: Tissue-specific biological roles of PPARγ ligand-dependent dysregulation in nonalcoholic fatty liver disease. *PPAR Res 2014*, 432647. doi:10.1155/2014/432647

Shoemaker, J. E., Gayen, K., Garcia-Reyero, N. et al. (2010). Fathead minnow steroidogenesis: In silico analyses reveals tradeoffs between nominal target efficacy and robustness to cross-talk. *BMC Syst Biol 4*, 89. doi:10.1186/1752-0509-4-89

Simon, T. W., Simons, S. S., Jr., Preston, R. J. et al. (2014). The use of mode of action information in risk assessment: Quantitative key events/dose-response framework for modeling the dose-response for key events. *Crit Rev Toxicol 44*, 17-43. doi:10.3109/10408444.2014.931925

Sonich-Mullin, C., Fielder, R., Wiltse, J. et al. (2001). IPCS

conceptual framework for evaluating a mode of action for chemical carcinogenesis. *Regul Toxicol Pharmacol 34*, 146-152. doi:10.1006/rtph.2001.1493

Suter, G. W., II and Cormier, S. M. (2011). Why and how to combine evidence in environmental assessments: Weighing evidence and building cases. *Sci Total Environ 409*, 1406-1417. doi:10.1016/j.scitotenv.2010.12.029

Thomas, R. S., Allen, B. C., Nong, A. et al. (2007). A method to integrate benchmark dose estimates with genomic data to assess the functional effects of chemical exposure. *Toxicol Sci 98*, 240-248. doi:10.1093/toxsci/kfm092

Tollefsen, K. E., Scholz, S. Cronin, M. T. et al. (2014). Applying adverse outcome pathways (AOPs) to support integrated approaches to testing and assessment (IATA). *Regul Toxicol Pharmacol 70*, 629-640. doi:10.1016/j.yrtph.2014.09.009

Tong, W., Perkins, R., Xing, L. et al. (1997). QSAR models for binding of estrogenic compounds to estrogen receptor alpha and beta subtypes. *Endocrinology 138*, 4022-4025. doi:10.1210/endo.138.9.5487

USEPA – US Environmental Protection Agency (2005). *Guidelines for Carcinogen Risk Assessment*. Risk Assessment Forum, US Environmental Protection Agency, Washington, DC.

van Wezel, A. P. and Opperhuizen, A. (1995). Narcosis due to environmental pollutants in aquatic organisms: Residue-based toxicity, mechanisms, and membrane burdens. *Crit Rev Toxicol 25*, 255-279. doi:10.3109/10408449509089890

Verhaar, H. J., Van Leeuwen, C. J. and Hermens, J. L. (1992). Classifying environmental pollutants. 1: Structure-activity relationships for prediction of aquatic toxicity. *Chemosphere 25*, 471-491. doi:10.1016/0045-6535(92)90280-5

Villeneuve, D. L., Mueller, N. D., Martinović, D. et al. (2009). Direct effects, compensation, and recovery in female fathead minnows exposed to a model aromatase inhibitor. *Environ Health Perspect 117*, 624-631. doi:10.1289/ehp.11891

Villeneuve, D. L., Breen, M., Bencic, D. C. et al. (2013). Developing predictive approaches to characterize adaptive responses of the reproductive endocrine axis to aromatase inhibition: I. Data generation in a small fish model. *Toxicol Sci 133*, 225-233. doi:10.1093/toxsci/kft068

Villeneuve, D. L., Crump, D., Garcia-Reyero, N. et al. (2014a). Adverse outcome pathway (AOP) development I: Strategies and principles. *Toxicol Sci 142*, 312-320. doi:10.1093/toxsci/kfu199

Villeneuve, D. L., Crump, D., Garcia-Reyero, N. et al. (2014b). Adverse outcome pathway development II: Best practices. *Toxicol Sci 142*, 321-330. doi:10.1093/toxsci/kfu200

Volz, D. C., Belanger, S. Embry, M. et al. (2011). Adverse outcome pathways during early fish development: A conceptual framework for identification of chemical screening and prioritization strategies. *Toxicol Sci 123*, 349-358. doi:10.1093/toxsci/kfr185

Weber, P., Medina-Oliva, G., Simon, C. and Iung, B. (2012). Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas. *Eng Appl Artif Intel 25*, 671-682. doi:10.1016/j.engappai.2010.06.002

Webster, A. F., Chepelev, N., Gagné, R. et al. (2015). Impact of genomics platform and statistical filtering on transcriptional benchmark doses (BMD) and multiple approaches for selection of chemical point of departure (PoD). *PLoS One 10*, e0136764. doi:10.1371/journal.pone.0136764

Weed, D. L. (2005). Weight of evidence, a review of concept and methods. *Risk Anal 25*, 1545-1557. doi:10.1111/j.1539-6924.2005.00699.x

Zang, Q., Rotroff, D. M. and Judson, R. S. (2013). Binary classification of a large collection of environmental chemicals from estrogen receptor assays by quantitative structure-activity relationship and machine learning methods. *J Chem Inf Model 53*, 3244-3261. doi:10.1021/ci400527b

Conflict of interest
The authors declare that they have no conflicts of interest.