CAUSAL, PREDICTIVE OR OBSERVATIONAL? DIFFERENT UNDERSTANDINGS OF KEY EVENT RELATIONSHIPS FOR ADVERSE OUTCOME PATHWAYS

Zheng Zhou[1,✉], Jeroen L.A. Pennings[2], and Ullrika Sahlin[1]

[1] Center for Environmental and Climate Science, Lund University, Sweden

[2] National Institute for Public Health and the Environment, the Netherlands

✉ Correspondence: Zheng Zhou <zheng.zhou@cec.lu.se>

1 INTRODUCTION

Regulatory toxicology and environmental risk assessment in the twenty-first century need to handle the assessment of ever-increasing numbers of chemicals, and at the same time adhere to the 3R principles to reduce, refine and replace animal experiments (Pallocca et al. 2022). The Adverse Outcome Pathways (AOPs) methodology facilitates the extrapolation of evidence at lower biological levels from animal and non-animal methods for the inference of adverse outcomes at higher biological levels (Villeneuve et al. 2014a). An AOP is defined as an evolving conceptual framework to organize existing knowledge pertaining to a molecular initiating event (MIE), an apical adverse outcome of interest (termed adverse outcome; AO), key measurable/ observable events (termed key events; KEs), and the linkages among MIE, AO and KEs (termed key event relationships; KERs) (Villeneuve et al. 2014a). The robustness and reliability of the extrapolation across events at different biological levels therefore depends on the characterization and inference of KERs (Villeneuve et al. 2024), which could be causal, observational and predictive.

Causal inference pursues the unbiased understanding of how one condition, the cause, affects another condition, the outcome, free from the threats to internal validity (Guzelian et al. 2005). Randomized experiments are the gold standard for causal inference (Guzelian et al. 2005; James et al. 2015) as they can approximate unbiased counterfactual estimates of treatment effects (Höfler 2005; Pearl 2009). Toxicological studies typically implement randomized experiments in controlled laboratory conditions, which reveals unbiased estimates of the association between the randomized assignment (e.g. dose) to the compound of interest and the experimental outcome (e.g. response), and thus causal

(Höfler 2005). The mechanism of randomized controlled trials is reviewed in the supplementary information.

When randomized experiments are not feasible, inference could rely on observational data. Quasi-experimental methods involve selecting subgroups from observational data to approximate unbiased average treatment effects, assuming specific conditions are met (Angrist and Pischke 2009). These methods have been employed in various fields such as economy, social science, psychology and epidemiology (Angrist and Pischke 2009; Bagiella et al. 2015; Fréchette et al. 2022). Observational data can also be evaluated using evidence-based approaches. The Bradford Hill considerations (Hill 1965), an early evidence-based framework for causal inference on observational, epidemiological data, has been modified to infer causal estimates in various contexts (Guzelian et al. 2005). For example, the evolved rank-ordered Hill considerations for mode of action/ human relevance framework can evaluate the weights/ degrees of evidence supporting hypothesized KERs in modes of action (MOAs) (Meek et al. 2014). However, neither quasi-experimental methods nor evidence-based frameworks fully guarantee unbiased estimates of causality. Therefore, results based on observational data have lower reliability compared to those from randomized experiments, which are termed as "causal estimates" in fields such as econometrics and social science. Within this article, we use the term "observational" to distinguish causal estimates from counterfactual causal findings, emphasizing the observational nature of the data.

Despite the involvement of experimental toxicological data, there could be challenges to obtain the empirical evidence needed for causal inference of KERs (Spînu et al. 2022). The

association between the dose and an outcome is causal, while the association between two outcomes (KEs) is not necessarily causal, due to the inability to randomize upstream KEs. For example, consider two KEs observed from the same dose-response experiment, with a hypothetical KER between them (Figure 1). This experiment generates two pairs of dose-response data, one for KE_{up} and one for KE_{down} respectively. While this setup provides unbiased estimates of the dose effects on KE_{up} and KE_{down} (Figure 1 left panel), KE_{up} itself is the consequence of the treatment and thus not randomized among subjects. The lack of randomization in KE_{up} implies that the estimates of the KER may not be causal (Figure 1 right panel). Direct evaluation of observed changes in KE_{down} conditional on KE_{up} is vulnerable to the threats to internal validity (Lipton and Ødegaard 2005), such as omitted variable bias from confounding. An instance of confounding bias is that, even if KE_{up} does not actually cause change in KE_{down}, simultaneous changes in both KEs might still occur due to a third, unobserved confounding event. The relationship between KE_{up}, KE_{down} and a potential confounder differs from the concept of non-adjacent KERs described in the handbook (Villeneuve et al. 2024). This is because a confounder might affect both KE_{up} and KE_{down} while a non-adjacent KER is when a third KE, which may not affect KE_{up}, is bridged over to KE_{down}. The estimated effect of KE_{up} on KE_{down} would be biased without isolating the confounding bias. For example, overdose of acetaminophen can cause hepatotoxicity (Hinson et al. 2010) and nephrotoxicity (Mazer and Perrone 2008). Both liver and kidney damage are suggested to be mediated by the depletion of glutathione (Jaeschke and McGill 2015; Mazer and Perrone 2008). The normal function of glutathione is involved in the detoxification of N-acetyl-p-benzoquinone imine, a toxic metabolite of acetaminophen (Casarett 2008). If the role of glutathione depletion is not properly accounted for, it could

introduce confounding bias to falsely suggest hepatotoxicity leads to nephrotoxicity upton acetaminophen overdose.

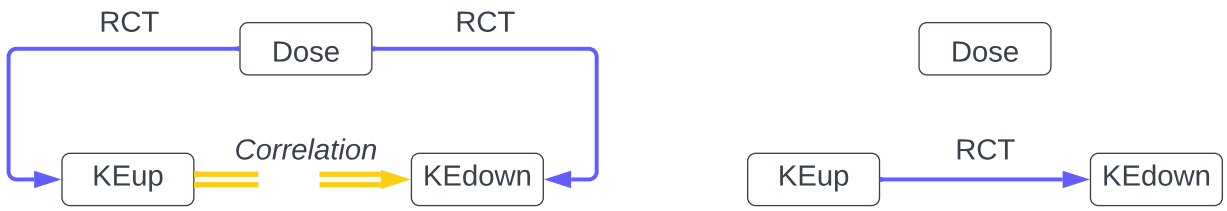

Figure 1: Directed acyclic graph of key event relationship evaluation. Left shows that treatment effects are causal on both key events respectively; right shows what is needed for a causal evaluation

Predictive models have been used in the application of traditional mechanistic study of toxicology (Helma 2005), such as physiologically based pharmacokinetic models (Sager et al. 2015). It also contributes to the development and application of alternatives to animal methods (Khabib et al. 2022), such as quantitative structure activity relationships (Barber et al. 2024; OECD 2014) and quantative system toxicology (Bloomingdale et al. 2017).

The difference between causal and predictive inference has led to debate and confusion (Young 2019). Causal inference falls under the framework of treatment effect evaluation (Young 2019), whose goal is to approximate unbiasedness (Young 2019) using randomized experiments and, when those are not possible, non-experimental methods (quasi-experimental, evidence-based frameworks) (Angrist and Pischke 2009). In contrast, the objective of predictive inference is to make accurate predictions. The term "predictive" term reflects the scope of utility and application, and falls outside the treatment effect framework (Young 2019). Predictions can be accurate even when biased by confounders

(Pearl 2009). For example, social-economic status is recognized as a major confounder to the adverse neurological effects of lead exposure during early childhood (Bellinger et al. 1992). However, some early studies accurately predicted neurological deficits based on dentine lead levels among children without adjusting for social-economic status (Herbert et al. 1979).

The difference between "causal" and "predictive" terms can be further understood in relation to mechanism and quantification. Mechanism implies that a causal relationship defines a directed linkage from the cause to the outcome, whereas the opposite is the error of reverse-causality (Russo and Williamson 2007). There is also a temporal requirement in a causal relationship that the cause must occur prior to the outcome, as emphasized through various iterations of Hill considerations (Becker et al. 2015; Cox 2018; Hill 1965; Meek et al. 2014). In contrast, a predictive relationship can be non-directional, i.e., knowing the temporal order between events is not necessary for prediction. Quantification implies that a predictive relationship is validated on a strong statistical strength of association. In contrast, a causal relationship does not necessarily include a description on the magnitude of causality. Two events with a causal relationship could be even statistically independent of each other, i.e., zero correlation. We present two examples in the supplementary information for these two points respectively.

Furthermore, practical differences between predictive and the treatment effect frameworks exist in many commonly used methods. Consider a linear regression with multiple covariates, whose parameters are optimized using maximum likelihood estimation. This regression can be used for both prediction and causal inference (Young

2019). When this regression is used as a predictive model, the goal is to improve the model's predictive performance(Young 2019), through the selection and optimization of the design matrix. When this regression is used as a causal model, bias adjustment is performed on the data prior to model construction, such as control selection (Cinelli et al. 2022) and backdoor elimination (Pearl 2009). The goal is to use the adjusted data and identify predictors that have significant coefficients (p values using maximum likelihood estimation and null hypothesis testing) (Angrist and Pischke 2009). The same model works differently for prediction and treatment effect evaluation. On the one hand, bias adjustment on data may not improve predictive performance (Brookhart et al. 2010; Davies et al. 2013). Including predictors of local significance may not improve global performance (Obermeyer and Emanuel 2016). On the other hand, optimizing the design matrix can not reduce the bias without prior bias adjustment (Chen et al. 2021).

The distinction between causal, observational and predictive inference is particularly relevant to context of toxicology and AOP development. This difference should be considered in the assembly of KEs and KERs. Traditionally, KEs and KERs are characterized among measured/ observed events from existing data, which is retrospective and evidence-based. In contrast, recent *in silico* predictive models enable prospective forecast of KEs and KERs for new compounds with minimal structural analysis (Hemmerich and Ecker 2020). The distinction between causal, observational and predictive inference should also be considered taking the difference between MOAs and AOPs. A major distinction between MOAs and AOPs is that MOAs are chemical specific while AOPs are not (Villeneuve et al. 2014a; Villeneuve et al. 2024). Evaluation of KERs under the MOA framework (Boobis et al. 2006) involves the evolved MOA Hill considerations (Meek et al. 2014), where the

confidence in KER causality is crucial along the MOA of the compound of interest. In the context of AOP development, it has been suggested that the causality is necessary between the AO and compounds on all AOP models including predictive ones (Spînu et al. 2022). However, it remains unclear whether it is necessary to require causality in all KERs in a chemical agnostic AOP. Given the chemical agnostic nature of AOPs, the supporting data for AOP development may include wide range of compounds, making it common to observe varying KE activities across compounds. Therefore, it is crucial to consider the appropriate application of causal, observational and predictive inference in AOP development.

Given the difference between the terms "causal", "observational" and "predictive" and its relevance to KER characterization, the objectives of this study are to: 1) examine the use of terminology describing KERs across AOP documentations and 2) investigate the adaptation of different terms in the development and application of AOPs. Achieving these objectives may facilitate a continuous development of AOP guidelines and a pragmatic implementation of AOP framework in academic, regulatory and industrial contexts.

2 METHODS

2.1 Terms for Key Event Relationships

We performed a text analysis on two AOP documentations focused on setting the definitions, principles and guidelines for AOP development: AOP developer's handbook (Villeneuve et al. 2024) (abbreviated as the handbook) and OECD Guidance Document for the Use of Adverse Outcome Pathways in Developing Integrated Approaches to Testing and Assessment (OECD 2017) (abbreviated as OECD guidance thereafter). Descriptions of KER

were located in the documents with the following keywords: “key event relationship” OR “KER” OR ((relationship OR link) AND “key event”).

2.2 Adaptation of terminology

We searched journal articles published by May 2024 to identify how different terminology has been adapted in development and application of AOPs. Our search was performed using online scholarly databases such as Google Scholar and PubMed. We gave preferences to sources from the “Animal-free Safety assessment of chemicals: Project cluster for Implementation of novel Strategies” (ASPIS) research cluster, including: ONTOX (ONTOX 2024), RISK-HUNT3R (RISK-HUNT3R 2024) and PrecisionTox (PrecisionTox 2024). We prioritized information from ASPIS due to the alignment between this article’s emphasis on AOPs and ASPIS’s perspective on non-animal methodologies. These searches aimed to identify exemplary cases to illustrate typical scenarios where the terms describing KERs have been adapted, rather than to conduct a systematic review of all relevant literature.

The identified information was organized by how the method of adaptation fits into the stages of AOP development and quantitative AOP (qAOP). AOP development consists of three operationally defined stages (OECD 2017; Villeneuve et al. 2014a). In contrast, there is no established guidance or consensus on the definition or stages of qAOP. Therefore, acting phases for qAOP application were based on pragmatic objectives from published literatures (Conolly et al. 2017; Perkins et al. 2019; Spinu et al. 2020), including prediction of AO (Jeong et al. 2018), derivation of point of departure (POD) (Perkins et al. 2019), and optimization of testing strategies (OECD 2017; Perkins et al. 2019).

3 RESULTS

3.1 Usage of terminology

KER descriptions in the OECD guidance are primarily based on the AOP developer's handbook, with minor editorial changes for the use in integrated testing strategies. Therefore, results of our text analysis focuses on KER descriptions identified from the handbook, highlighting the text examples located from the sections where the definitions and principles of AOP development are set. Results show that with a clear distinction between the terms "causal" and "predictive", the handbook's usage of them varies across sections (Table 1). For instance, in the handbook Section 3 KER Description on page 25, KERs are described in different ways: as predictive *"Within the AOP framework, the predictive relationships that facilitate extrapolation are represented by the KERs"* or as causal *"KERs are always described in the form of a directed relationship (one-way arrow) linking an upstream "causing" event to a downstream "responding" event"*. In some instances KERs are described with both "causal" and "predictive", e.g., on page 6 Table 1 Definitions of Key Terms that *"A scientifically-based relationship that connects one key event to another, defines a causal and predictive relationship between the upstream and downstream event"*. Compared to instances where "causal" or "predictive" are used separately, the inclusion of both terms suggests the consideration of dual requirements in KER description. The dual requirements does not necessarily imply a conflict against separate use of terms. When both causal and predictive inference are needed for KER description, one or both may be emphasized at different stages of AOP development.

The dual requirements are also evident in the handbook's protocol for evaluating KER supporting evidence. The handbook evaluates KER supporting evidence (Section 3G) through 1) biological plausibility , 2) empirical evidence and 3) current quantitative understandings. The handbook's guidance to assess biological plausibility (annex 1) is adapted from the evolved MOA Hill considerations (Meek et al. 2014), which is observational as mentioned in the Introduction of this paper. Empirical evidence is assessed (handbook page 31 subsection ii) for predictive utility that *"assessment of changes in the upstream KE (KEupstream) leading to, or being associated with, a predictable subsequent change in the downstream KE (KEdownstream)"*. Moreover, within the same section, the uncertainty and inconsistency of empirical evidence is evaluated that *"while there are expected patterns of concordance that support a causal linkage between the KEs in the pair, it is also helpful to identify experimental details that may explain apparent deviations from the expected patterns of concordance"*. The handbook's evaluation of current quantitative understandings of KERs, considering the weights of empiricial evidence supporting KER causality, is dedicated to prediction *"to define the precision with which the state of the downstream KE can be predicted from knowledge of the state of the upstream KE"* (page 34). Therefore, it seems that the handbook sets the definitions, descriptions and protocols for evaluation of KERs based on a dual focus of causal and predictive inference.

The term "observational" or the more commonly recognized "causal estimates" does not appear in the handbook, nor the OECD guidance.

Table 1: Occurrence of terminology describing KERs in two AOP documentations

		Terminology	
Location of KER description	Quote	Causal	Predictive
OECD guidance document	***OECD guidance document 2017***	**OECD guidance document**	
AOP definition	*a sequential chain of causally linked events at different levels of biological organisation*	√	
	AOP handbook 2024 version 2.7		
Definitions of key terms, KER , page 6 Table 1	*A scientifically-based relationship that connects one key event to another, defines a causal and predictive relationship between the upstream and downstream event*	√	√
Section 1, AOP development tip #3, page 13 line 448	*an AOP is acceptable when there are multiple KEs, causally linked to the MIE and AO*	√	
Section 3 KER Description, page 25 line 929	*Within the AOP framework, the predictive relationships that facilitate extrapolation are represented by the KERs.*		√
Section 3 KER Description, page 25 line 941	*KERs are always described in the form of a directed relationship (one-way arrow) linking an upstream "causing" event to a downstream "responding" event.*	√	
Section 3G Evidence Supporting this KER, Empirical Evidence, page 29 line 1126	*biological plausibility, empirical support, and the current quantitative understanding of the KER are evaluated with regard to the predictive relationships/associations between defined pairs of KEs as a basis for considering WoE (Section 4)*		√
Section 3G Evidence Supporting this KER, Empirical Evidence, page 31 line 1167	*evidence relevant to assessment of changes in the upstream KE (KEupstream) leading to, or being associated with, a predictable subsequent change in the downstream KE (KEdownstream).*		√
Section 3G Evidence Supporting this KER,	*while there are expected patterns of concordance that support a causal linkage between the KEs in the pair, it is also helpful to identify experimental details*	√	

		Terminology	
Location of KER description	Quote	Causal	Predictive
Uncertainties and Inconsistencies, page 32 line 1242	*that may explain apparent deviations from the expected patterns of concordance.*		
Section 4B Assess the Essentiality of ALL KEs, Direct evidence, page 39, line 1470	*evidence that overexpression of the object of the KE prevents or impacts the downstream KEs in a manner consistent with its causal, and essential, role in the pathway.*	√	
Section 4C Evidence Assessment, Review the Empirical Support for Each KER, page 41, line 1565	*Empirical support entails consideration of experimental data in terms of the associations between KEs – namely dose-response concordance and temporal relationships between and across multiple KEs.*		√

3.2 Adaptation of terminology in AOP development and qAOP application

Table 2: Examples of adaptation of KER terminology in AOPs and qAOPs from published literature

Stage	Characteristics	Causal	Observational	Predictive
Putative AOP development	Assembly of hypothesized KEs and KERs	Through targeted blocking experiment (Skarnes 2011)	Through WoE supporting KE essentiality and KER biological plausibility (Becker et al. 2015; Jensen et al. 2023)	Through statistical inference: network analysis (Villeneuve et al. 2018) and machine learning (Gadaleta et al. 2022)
	Assembly of partial AOP with incomplete linkage		As a result of known gaps and uncertainties (Bal-Price et al. 2015; Becker et al. 2015)	

Stage	Characteristics	Causal	Observational	Predictive
	between MIE and AO			
Qualitative AOP development	Assembly of KEs		Through describing possibility of measurement (Becker et al. 2015)	
	Assembly of KERs		Through describing WoE of empirical concordance supporting KERs (Becker et al. 2015); Through evaluation of KERs using quasi-experimental methods (Landfors et al. 2023)	
	Evaluation of AOP		Through overall WoE supporting the AOP: biological plausibility, essentiality and empirical evidence (Becker et al. 2015)	
Quantitative AOP development	Assembly of KEs		Through describing accuracy and precision of measurements (Becker et al. 2015)	
	Assembly of KERs		Through evaluating WoE for characterizing the magnitude and/or duration of change in upstream KE needed to evoke change in downstream KE (Becker et al. 2015)	Same as in observational but with a focus on predictive performance of response-response models (Becker et al. 2015); Through characterizing known gaps and uncertainties between MIE and AO: known modulating factors (Becker et al. 2015)
Quantitative AOP application (acting)	Prediction of AO			Based on KEs of good accuracy and precision of measurement (Jeong et al. 2018)

Stage	Characteristics	Causal	Observational	Predictive
	Derivation of POD			For gene expression/ transcription (Song et al. 2023) and AO (Perkins et al. 2019)
	Optimization of testing strategy			Through integrated approaches to testing and assessment (OECD 2017)

KE: key event; KER: key event relationship; AO: adverse outcome; MIE: molecular initiating event; AOP: adverse outcome pathway; POD: point of departure; WoE: weight of evidence

Empty cell indicates no example identified for the corresponding adaptation

The handbook's dual requirement on both causal and predictive KERs provides a comprehensive framework for AOP development. Examples identified from published literature are categorized by which term among "causal", "observational" and "predictive" was adapted, and displayed throughout stages (Villeneuve et al. 2014a) of putative, qualitative and quantitative AOP development (Table 2).

Adapting the term "causal" in the strict sense would require experiments implementing direct randomization of upstream KEs. However, no examples of practical implementation could be identified in the literature. The alternative approach in published studies is targeted blocking experiments to infer the impact of KE_{up} on KE_{down} that a reduction or complete absence of a KE_{down} when KE_{up} is blocked (Meek et al. 2014). Empirical evidence from targeted blocking experiments are direct support and contribute to a high level of confidence for evaluating the essentiality of KEs (Meek et al. 2014; Villeneuve et al. 2014b; Villeneuve et al. 2024). The development and application of qAOPs often assumes that KEs

are essential (Jensen et al. 2023) and KERs are causal (Burgoon et al. 2020; Gou et al. 2023).

Observational analyses of KEs, KERs and overall AOP could involve weight-of-evidence (WoE) assessments of supporting evidence and quasi-experimental assessments. The evolved MOA Hill considerations (Meek et al. 2014) are adapted in the handbook and the OECD guidance for the WoE assessment of the essentiality of KEs , the biological plausibility of KERs, the concordance of KER empirical evidence and the overall AOP level of confidence (OECD 2017; Villeneuve et al. 2024). Best practice has been recommended for the adapted WoE assessments (Villeneuve et al. 2014b) to categorize confidence levels as high, moderate or low (Becker et al. 2015). Characterizing known gaps and uncertainties that modulate incomplete linkages between MIE and AO could further enhance the weights of evidence in the overall AOP (Bal-Price et al. 2015; Villeneuve et al. 2024). When high WoE is achieved, qualitative and quantitative descriptions of KE measurement methods and their accuracy and precision are crucial (Villeneuve et al. 2014b). These descriptions support the characterization and calibration of predictive modeling of relationships among the sequence of KEs in the overall AOP (Becker et al. 2015). In addition to WoE assessments, the instrumental variable method, a type of quasi-experimental design, shows promise in providing strong evidence supporting KER causality (Landfors et al. 2023).

Predictive models are primarily involved in the putative and quantitative stages (OECD 2017). Assembly of KEs and KERs in the putative stage can be derived using machine learning (Gou et al. 2023; Hemmerich and Ecker 2020) or probabilistic graphical network models (Villeneuve et al. 2018). KEs could be screened from a pool of candidate molecular

and cellular events using feature selection techniques (Perkins et al. 2019). Similar methods have been applied to screen for genes and genomic modules associated with significant differential gene expressions (Callegaro et al. 2021; Hemmerich and Ecker 2020). Topological metrics are evaluated to identify nodes and edges that best improve overall predictive performance on the AO as KEs and KERs, respectively (Villeneuve et al. 2018). Graphical network models might consist of deterministic and probabilistic nodes, and undirected, directed, acyclic edges, that have been involved in existing studies of AOP development and application (OECD 2017; Perkins et al. 2019; Villeneuve et al. 2018). Quantitative AOP development focuses on the derivation of response-response models [Villeneuve2024]. Studies of qAOP typically assume the causality of KERs has been established (Burgoon et al. 2020; Gou et al. 2023; Jensen et al. 2023), supported by the essentiality of KEs, the biological plausibility of KERs and the concordance in KER empirical evidence. Furthermore, to assure the reliability of qAOPs within relevant applicability domains, the handbook includes a guidance (annex 2) to evaluate the WoE of qAOPs by not only predictive performance, also considerations for known modulating factors, feedback loops and generalizability across the applicability domains. Examples of possible regulatory application of qAOPs include prediction of AO given MIE (Jeong et al. 2018), derivation of omics (Song et al. 2023) and *in vitro* (Perkins et al. 2019) PODs, and optimization of testing strategies (OECD 2017).

In summary, the handbook's nuanced usage of terms are reflected in AOP development that focus of adaptation varies across stages of AOP development. Given the challenge of obtaining experimental evidence for achieve causality, targeted block experiments offer a practical approach to balance weights of evidence and feasibility. Best practice (Becker et

al. 2015; Villeneuve et al. 2014b) has been recommended to implement observational inference using WoE assessments based on the evolved MOA Hill considerations (Meek et al. 2014). The stage of putative AOP development has a dual focus on observational and predictive evaluations. The qualitative stage focuses on the WoE of evidence supporting KEs, KERs and the overall AOP. The stage of qAOP application and application involves various predictive approaches, including machine learning, network and response-response models.

4 DISCUSSIONS

The nuanced terminology usage in the handbook and approaches of adaptation reflect the complex nature of AOP development. Our results suggest several issues that may warrant further discussion, including: 1) a mismatch between the terminology in guideline and the methodology in practice, 2) considerations of terminology for interdisciplinary communication and

4.1 Mismatch between methodology and terminology

Some putative AOP studies inferred causality of KERs based on predictive models (Gou et al. 2023; Jeong et al. 2018). We argue the use of predictive models in these studies supported the predictive but not the causal aspect of KERs, highlighting a mismatch between the treatment effect framework and predictive models. This mismatch could be demonstrated through possible misinterpretations of probabilistic network models as causal. Probabilistic network models (Koller and Friedman 2009) are based on directed acyclic graphs (DAGs). DAGs are commonly involved in causal reasoning and designing of

structural equation modeling (Greenland et al. 1999), e.g. to approximate unbiasedness by quasi-experimental methods. But the use of a DAG is not a guarantee for having a causal model. It depends on context, if a model involving DAGs can provide unbiased average treatment effects for causal inference. In Bayesian modelling, Directed acyclic graphs are used to decompose the joint probability distribution of variables and parameters into a product, with no implication of causality. A Bayesian network is usually referred to a probabilistic network model with binary or categorical nodes. Bayesian networks comprise a wide range of applications, where the network can be inferred from multivariate data, e.g. in machine learning, or given as representative of causal relationships where the probabilities or conditional probability distribution are informed by expert judgment, e.g. in decision analysis. Bayesian networks have been applied to parameterize models where confounding biases are adjusted using quasi-experimental methods such as the the backdoor criterion (Pearl 2009) and instrumental variables (Shingaki et al. 2021). With the popularity of Bayesian networks (Kaikkonen et al. 2021; Spinu et al. 2020) comes misconceptions of to what extent a Bayesian network indicates causality, which might come from the use of a DAG. There are studies where results of a Bayesian network were interpreted as causal without employing bias reduction (Chen and Pollino 2012; Jeong et al. 2018; Zou et al. 2010). For evaluating the uncertainty in a Bayesian network, beyond the inherent uncertainty embedded within the probability model, it is crucial to consider the process and context in which the Bayesian network is applied (Sahlin et al. 2021).

This mismatch is particularly crucial in the stage of putative AOP development, where the assembly of KEs and KERs could be achieved through either causal inference or statistical modeling. The reason is, although predictive inference of KEs and KERs is optional in the

recommended best practice for the putative stage (Villeneuve et al. 2014b), an AOP would be incomplete without considerations on the essentiality of KEs, biological plausibility and empirical support of KERs. Therefore, misrepresenting results from predictive models as causal may falsely imply that the dual requirements have been fulfilled in the putative stage, which could risk overlooking causal and observational assessment in the qualitative stage.

Furthermore, this mismatch also highlights the importance to integrate predictive models with the evidence from causal or observational inference in the qAOP stage. Current qAOP development often assumes that causality of KERs is well-supported in established AOPs (Burgoon et al. 2020; Gou et al. 2023; Jensen et al. 2023). Yet there are limited approaches to propagate the assumed causality in predictive models. One potential approach is the use of Bayesian methods to integrate known causality of KERs and prior knowledge on KEs in qAOP (Spînu et al. 2022). Another approach involves incorporating causal KERs on the target/ utility/ loss function of predictive models through regularization techniques such as Lasso (Tibshirani 1996). While regularization has been applied in in silico model to mitigate overfitting (Hemmerich and Ecker 2020), there has been limited exploration of its use in adjusting the target/ utility/ loss function on known causality. This may be due to concerns that regularizing on causality may penalize predictive performance. However, it is crucial to re-emphasize that AOPs are only complete with both causal and predictive characteristics (Becker et al. 2015) and good predictive performance alone is insufficient for robust qAOP development. According to the handbook's guideline on the WoE assessment of qAOPs (annex 2), known modulating factors and feedback loops need to be

considered along with predictive performance for a qAOP to achieve high level of confidence (Villeneuve et al. 2024).

4.2 Careful considerations of terminology in AOP guidance

We would like to initiate a discussion for careful considerations on the usage of terms "causality" and "essentiality" to facilitate interdisciplinary communication and collaboration.

The use of term "causal" needs to be considered because the frameworks for evaluating causality in the handbook (Villeneuve et al. 2024) and for the use in integrated testing strategies (OECD 2017) may not necessarily conclude causality in the same sense employed in toxicology and generic scientific reasoning. Even the recommended best practice (Becker et al. 2015; Villeneuve et al. 2014b) may be insufficient to achieve unbiased estimates of KE essentiality and KER causality. This is not a critique of current established AOP guidelines and practice because, as shown by our results, the term "causal" is pronounced and utilized in a consistent manner by the handbook, the OECD guidance and studies within the niche of AOP development. Distinctive usage across fields of the same terminology is common. However, adding clarifying annotations on the term "causal" could greatly facilitate interdisciplinary communication and collaboration that potential bias may exist and that the results could be "causal estimates" rather than "causal".

To demonstrate, the handbook's framework highlights the inclusion of direct and indirect evidence and a prioritization of direct evidence in the assessment of essentiality of KEs and biological plausibility of KERs. Direct evidence of KE essentiality (handbook section 4B, Table 5) include evidence from targeted blocking experiments, such as know-out

experiments (Skarnes et al. 2011). Knockout experiments pragmatically approximates counterfactual inference and are more robust and reliable than observational evaluation of multiple KEs (see supplementary information). However, results from targeted blocking experiments are not always causal because the caveat of off-target issues applies in design (Höijer et al. 2022). The instrumentation of blocking could be attributable to confounding and uncontrolled mediation on downstream KEs (Skarnes et al. 2011). Unbiased causality could be achieved only if the blocking methods induce blocking in the target KE_{up} but not anything else that causes changes in KE_{down} or their modulating factors. If the instrument to block KE_{up} activates a third, unobserved event that, e.g., increases the activity of KE_{down}, it is possible to observe no change in KE_{down}, because the reduction by blocking KE_{up} could be canceled out by the activation from the third event.

In targeted knockout experiments, there could be at least two technical reasons leading to the off-target issue. Firstly, the genetic target in knockout procedure is identified by genome-wide association studies, which is association by definition. Confounders may exist since the gene-protein associations are not always one-to-one matched (Olariu et al. 2021). It is well known that the same protein can be coded by different genes or multiple copies of the same gene at different loci (Campenhout et al. 2019). It is also not uncommon to observe the synthesis of multiple proteins from the transcription of one gene. The variational splicing of introns of one gene can be transcripted to different mRNAs, which result in the production of multiple, distinctive proteins (De Conti et al. 2013). Targeted mutation techniques have the potential to become quasi-experimental by implementing bias adjustment methods such as instrumental variables (Landfors et al. 2023). Secondly, assuming the target identification is free from confounding bias, unintended genetic

changes could be induced during gene editing (Höijer et al. 2022). Therefore, direct evidence from targeted blocking experiments, as prescribed in the best practice, could be insufficient to guarantee causality.

Indirect, non-experimental evidence is evaluated using a WoE framework based on the evolved MOA Hill considerations (Meek et al. 2014). The evolved MOA Hill considerations classifies weights/ degrees of confidence into high, moderate or low rather than a dichotomous judgement of causality. This is evident that the terms "causal", "causality" or etc. were included only once (not including quotes) in the source paper (Meek et al. 2014) to describe the consideration for KE essentiality that *"if following cessation of repeated exposure for vaiours periods, effects are reverse, this constitutes relatively strong evidence that key events are causal"*. Furthermore, the adaptations made by the handbook implement no further bias adjustment and focus on the classification of WoE of supporting evidence.

In conclusion, the frameworks for KER evaluation in the handbook and recommended practice could be insufficient to reach the causal requirement prescribed in the handbook.

We also would like to bring up a discussion on the use of the term "essential", particularly in the essentiality of KEs. We argue that adding enhancing annotations on the term "essential" could benefit its practical utility in the assessment of KE essentiality. It is crucial to note that this discussion does not revolve around the distinction between mathematical and toxicological context of "essential". The term "essential" has specialized mathematical context and usage to describe the conditional relationship between two statements. To claim mathematical essentiality of event A for event B, that is to say, if A does not occur, then B cannot occur at all. The validation of mathematical essentiality requires a falsity test.

Any case where the occurrence of B without the occurrence of A falsifies essentiality. Obviously, such a falsity test is practically difficult to perform in science (Graesser and Hu 2011), especially in biology and toxicology. Most biological events are the consequence of one or several alternative sufficient causes. The use of "essential" in toxicology is distinct from the mathematical context, including the evolved MOA Hill considerations and AOP development. Under the toxicological context, evaluation of essentiality considers the impact of blocking or reversing KE_{up} on KE_{down} (Meek et al. 2014). A mathematically essential event would cause complete absence in the consequence, while a toxicologically essential event could describe a KE_{up} which, when blocked, prevent or significantly modify KE_{down} (Meek et al. 2014; Villeneuve et al. 2024). The validation of mathematical essentiality of KE_{up} for KE_{down} could be practically unfeasible. The guidelines for AOP development needs to be a pragmatic framework that balances scientific rigor and practicability to serve regulatory and industrial application in the real world. However, the level of significance of modification/ reduction is subjective. Neither the evolved MOA Hill considerations nor the handbook defines the level of modification in KE_{down} to which the changes associated with blocking KE_{up} should be considered as significant and sufficient. Furthermore, the level of modification/ reduction is diverse across systems, organs and molecules by nature. For example, body weight change depends on a myriad of factors, including time period, age, sex and baseline weight. It would be helpful to annotate the handbook's defining question of KE essentiality by adding a description of the level at which the modification in KE_{down} is both statistically significant and biologically sufficient. Further research using the benchmark dose modeling methodology might be useful to characterize the level of modification for KE essentiality.

5 CONCLUSIONS

This study aimed to explore the usage and adaptation of the terms "causal", "observational" and "predictive" for KER description in AOP guidance documents. The results indicate that KERs are described to fulfill both causal and predictive requirements across three stages of AOP development, and distinct approaches are implemented to address different focuses of adaptation across these stages. These findings suggest that the nuanced approach towards the distinction between causal and predictive inference of KERs reveal a mismatch between predictive models and causality in published studies of AOP development, particularly pertaining to the putative stage. Further exploration is needed to integrate causality in qAOP development and application. Further discussions could be helpful to consider the handbook's usage of the terms "causal" and "essential" to facilitate more comprehensive, transparent interdisciplinary communication and practical application of AOP under the academic, regulatory and industrial contexts.

REFERENCE

Angrist JD, Pischke J-S. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Bagiella E, Karamlou T, Chang H, Spivack J. 2015. Instrumental variable methods in clinical research. The Journal of Thoracic and Cardiovascular Surgery 150:779–782; doi:10.1016/j.jtcvs.2015.07.056.

Bal-Price A, Crofton KM, Sachana M, Shafer TJ, Behl M, Forsby A, et al. 2015. Putative adverse outcome pathways relevant to neurotoxicity. Critical reviews in toxicology 45:83–91; doi:10.3109/10408444.2014.981331.

Barber C, Heghes C, Johnston L. 2024. A framework to support the application of the OECD guidance documents on (q)SAR model validation and prediction assessment for regulatory decisions. Computational Toxicology 100305; doi:10.1016/j.comtox.2024.100305.

Becker RA, Ankley GT, Edwards SW, Kennedy SW, Linkov I, Meek B, et al. 2015. Increasing scientific confidence in adverse outcome pathways: Application of tailored bradford-hill considerations for evaluating weight of evidence. Regulatory Toxicology and Pharmacology 72:514–537; doi:10.1016/j.yrtph.2015.04.004.

Bellinger DC, Stiles KM, Needleman HL. 1992. Low-level lead exposure, intelligence and academic achievement: A long-term follow-up study. Pediatrics 90:855–861; doi:10.1542/peds.90.6.855.

Bloomingdale P, Housand C, Apgar JF, Millard BL, Mager DE, Burke JM, et al. 2017. Quantitative systems toxicology. Translational Toxicology: Biomarkers 4:79–87; doi:10.1016/j.cotox.2017.07.003.

Boobis AR, Cohen SM, Dellarco V, McGregor D, Meek ME, Vickers C, et al. 2006. IPCS framework for analyzing the relevance of a cancer mode of action for humans. Critical Reviews in Toxicology 36:781–92; doi:10.1080/10408440600977677.

Brookhart MA, Rassen JA, Schneeweiss S. 2010. Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidemiology and Drug Safety 19:537–554; doi:10.1002/pds.1908.

Burgoon LD, Angrish M, Garcia-Reyero N, Pollesch N, Zupanic A, Perkins E. 2020. Predicting the probability that a chemical causes steatosis using adverse outcome pathway bayesian networks (AOPBNs). Risk Analysis 40:512–523; doi:10.1111/risa.13423.

Callegaro G, Kunnen SJ, Trairatphisan P, Grosdidier S, Niemeijer M, Hollander W den, et al. 2021. The human hepatocyte TXG-MAPr: Gene co-expression network modules to support mechanism-based risk assessment. Arch Toxicol 95:3745–3775; doi:10.1007/s00204-021-03141-w.

Campenhout CV, Cabochette P, Veillard A-C, Laczik M, Zelisko-Schmidt A, Sabatel C, et al. 2019. Guidelines for optimized gene knockout using CRISPR/Cas9. BioTechniques 66:295–302; doi:10.2144/btn-2018-0187.

Casarett LJ. 2008. *Casarett and doull's toxicology: The basic science of poisons*. McGraw-Hill New York.

Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. 2021. Ethical machine learning in healthcare. Annual Review of Biomedical Data Science 4:123–144; doi:https://doi.org/10.1146/annurev-biodatasci-092820-114757.

Chen SH, Pollino CA. 2012. Good practice in bayesian network modelling. Environmental Modelling & Software 37: 134–145.

Cinelli C, Forney A, Pearl J. 2022. A crash course in good and bad controls. Sociological Methods & Research 00491241221099552; doi:10.1177/00491241221099552.

Conolly RB, Ankley GT, Cheng W, Mayo ML, Miller DH, Perkins EJ, et al. 2017. Quantitative adverse outcome pathways and their application to predictive toxicology. Environmental Science & Technology 51:4661–4672; doi:10.1021/acs.est.6b06230.

Cox LAJr. 2018. Modernizing the bradford hill criteria for assessing causal relationships in observational data. Critical Reviews in Toxicology 48:682–712; doi:10.1080/10408444.2018.1518404.

Davies NM, Smith GD, Windmeijer F, Martin RM. 2013. Issues in the reporting and conduct of instrumental variable studies: A systematic review. Epidemiology 24.

De Conti L, Baralle M, Buratti E. 2013. Exon and intron definition in pre-mRNA splicing. WIREs RNA 4:49–60; doi:10.1002/wrna.1140.

Fréchette GR, Sarnoff K, Yariv L. 2022. Experimental economics: Past and future. Annual Review of Economics 14:777–794; doi:https://doi.org/10.1146/annurev-economics-081621-124424.

Gadaleta D, Spînu N, Roncaglioni A, Cronin MTD, Benfenati E. 2022. Prediction of the neurotoxic potential of chemicals based on modelling of molecular initiating events upstream of the adverse outcome pathways of (developmental) neurotoxicity. International Journal of Molecular Sciences 23; doi:10.3390/ijms23063053.

Gou X, Ma C, Ji H, Yan L, Wang P, Wang Z, et al. 2023. Prediction of zebrafish embryonic developmental toxicity by integrating omics with adverse outcome pathway. Journal of Hazardous Materials 448:130958; doi:10.1016/j.jhazmat.2023.130958.

Graesser AC, Hu X. 2011. Commentary on causal prescriptive statements. Educational Psychology Review 23:279–285; doi:10.1007/s10648-011-9163-x.

Greenland S, Pearl J, Robins JM. 1999. Causal diagrams for epidemiologic research. Epidemiology 37–48.

Guzelian PS, Victoroff MS, Halmes NC, James RC, Guzelian CP. 2005. Evidence-based toxicology: A comprehensive framework for causation. Human & Experimental Toxicology 24:161–201; doi:10.1191/0960327105ht517oa.

Helma C. 2005. *Predictive toxicology*. CRC Press.

Hemmerich J, Ecker GF. 2020. In silico toxicology: From structure–activity relationships towards deep learning and adverse outcome pathways. WIREs Computational Molecular Science 10:e1475; doi:10.1002/wcms.1475.

Herbert NL, Charles G, Alan L, Robert R, Henry P, Cornelius M, et al. 1979. Deficits in psychologic and classroom performance of children with elevated dentine lead levels. New England Journal of Medicine 300:689–695; doi:10.1056/NEJM197903293001301.

Hill AB. 1965. The environment and disease: Association or causation? Proceedings of the Royal Society of Medicine 58:295–300; doi:10.1177/003591576505800503.

Hinson JA, Roberts DW, James LP. 2010. Mechanisms of acetaminophen-induced liver necrosis. In: *Adverse drug reactions* (J. Uetrecht, ed). Springer Berlin Heidelberg:Berlin, Heidelberg. 369–405.

Höfler M. 2005. Causal inference based on counterfactuals. BMC Medical Research Methodology 5:28; doi:10.1186/1471-2288-5-28.

Höijer I, Emmanouilidou A, Östlund R, Schendel R van, Bozorgpana S, Tijsterman M, et al. 2022. CRISPR-Cas9 induces large structural variants at on-target and off-target sites in vivo that segregate across generations. Nature Communications 13:627; doi:10.1038/s41467-022-28244-5.

Jaeschke H, McGill MR. 2015. Cytochrome P450-derived versus mitochondrial oxidant stress in acetaminophen hepatotoxicity. Toxicology Letters 235:216–217; doi:10.1016/j.toxlet.2015.04.002.

James R, Britt J, Halmes N, Guzelian P. 2015. Evidence-based causation in toxicology: A 10-year retrospective. Human & Experimental Toxicology 34:1245–1252; doi:10.1177/0960327115601767.

Jensen MA, Blatz DJ, LaLone CA. 2023. Defining the biologically plausible taxonomic domain of applicability of an adverse outcome pathway: A case study linking nicotinic acetylcholine receptor activation to colony death. Environmental Toxicology and Chemistry 42:71–87; doi:10.1002/etc.5501.

Jeong J, Song T, Chatterjee N, Choi I, Cha YK, Choi J. 2018. Developing adverse outcome pathways on silver nanoparticle-induced reproductive toxicity via oxidative stress in the

nematode caenorhabditis elegans using a bayesian network model. Nanotoxicology 12:1182–1197; doi:10.1080/17435390.2018.1529835.

Kaikkonen L, Parviainen T, Rahikainen M, Uusitalo L, Lehikoinen A. 2021. Bayesian networks in environmental risk assessment: A review. Integrated Environmental Assessment and Management 17:62–78; doi:10.1002/ieam.4332.

Khabib MNH, Sivasanku Y, Lee HB, Kumar S, Kue CS. 2022. Alternative animal models in predictive toxicology. Toxicology 465:153053; doi:10.1016/j.tox.2021.153053.

Koller D, Friedman N. 2009. *Probabilistic graphical models: Principles and techniques*. MIT press.

Landfors F, Chorell E, Kersten S. 2023. Genetic mimicry analysis reveals the specific lipases targeted by the ANGPTL3-ANGPTL8 complex and ANGPTL4. Journal of Lipid Research 64; doi:10.1016/j.jlr.2022.100313.

Lipton R, Ødegaard T. 2005. Causal thinking and causal language in epidemiology: It's in the details. Epidemiological Perspectives and Innovations 2: 8–9.

Mazer M, Perrone J. 2008. Acetaminophen-induced nephrotoxicity: Pathophysiology, clinical manifestations, and management. Journal of Medical Toxicology 4:2–6; doi:10.1007/BF03160941.

Meek M, Palermo CM, Bachman AN, North CM, Jeffrey Lewis R. 2014. Mode of action human relevance (species concordance) framework: Evolution of the bradford hill considerations and comparative analysis of weight of evidence. Journal of Applied Toxicology 34:595–606; doi:10.1002/jat.2984.

Obermeyer Z, Emanuel EJ. 2016. Predicting the future - big data, machine learning, and clinical medicine. The New England journal of medicine 375:1216–1219; doi:10.1056/NEJMp1606181.

OECD. 2017. Guidance document for the use of adverse outcome pathways in developing integrated approaches to testing and assessment (IATA).

OECD. 2014. *Guidance document on the validation of (quantitative) structure-activity relationship [(q)SAR] models*.

Olariu V, Yui MA, Krupinski P, Zhou W, Deichmann J, Andersson E, et al. 2021. Multi-scale dynamical modeling of t cell development from an early thymic progenitor state to lineage commitment. Cell Reports 34:108622; doi:10.1016/j.celrep.2020.108622.

ONTOX. 2024. Scientific papers-ONTOX. ONTOX-project.

Pallocca G, Moné MJ, Kamp H, Luijten M, Van de Water B, Leist M. 2022. Next-generation risk assessment of chemicals - rolling out a human-centric testing strategy to drive 3R implementation : The RISK-HUNT3R project perspective. Alternatives to Animal Experimentation : ALTEX 39:419–426; doi:10.14573/altex.2204051.

Pearl J. 2009. Causal inference in statistics: An overview. Statistics Surveys 3:96–146; doi:10.1214/09-SS057.

Perkins EJ, Ashauer R, Burgoon L, Conolly R, Landesmann B, Mackay C, et al. 2019. Building and applying quantitative adverse outcome pathway models for chemical hazard and risk assessment. Environmental Toxicology and Chemistry 38:1850–1865; doi:10.1002/etc.4505.

PrecisionTox. 2024. Publications-PrecisionTox.

RISK-HUNT3R. 2024. Publications-RISK-HUNT3R. RISK-HUNT3R.

Russo F, Williamson J. 2007. Interpreting causality in the health sciences. International Studies in the Philosophy of Science 21:157–170; doi:10.1080/02698590701498084.

Sager JE, Yu J, Ragueneau-Majlessi I, Isoherranen N. 2015. Physiologically based pharmacokinetic (PBPK) modeling and simulation approaches: A systematic review of published models, applications, and model verification. Drug Metabolism and Disposition 43:1823; doi:10.1124/dmd.115.065920.

Sahlin U, Helle I, Perepolkin D. 2021. "This is what we don't know": Treating epistemic uncertainty in bayesian networks for risk assessment. Integrated Environmental Assessment and Management 17:221–232; doi:10.1002/ieam.4367.

Shingaki R, Kanda H, Kuroki M. 2021. Selection and integration of generalized instrumental variables for estimating total effects. Statistical Papers 62:2355–2381; doi:10.1007/s00362-020-01190-4.

Skarnes WC, Rosen B, West AP, Koutsourakis M, Bushell W, Iyer V, et al. 2011. A conditional knockout resource for the genome-wide study of mouse gene function. Nature 474:337–342; doi:10.1038/nature10163.

Song Y, Zheng K, Brede DA, Gomes T, Xie L, Kassaye Y, et al. 2023. Multiomics point of departure (moPOD) modeling supports an adverse outcome pathway network for ionizing radiation. Environmental Science & Technology 57:3198–3205; doi:10.1021/acs.est.2c04917.

Spinu N, Cronin MTD, Enoch SJ, Madden JC, Worth AP. 2020. Quantitative adverse outcome pathway (qAOP) models for toxicity prediction. Archives of Toxicology 94:1497–1510; doi:10.1007/s00204-020-02774-7.

Spînu N, Cronin MTD, Madden JC, Worth AP. 2022. A matter of trust: Learning lessons about causality will make qAOPs credible. Computational toxicology (Amsterdam, Netherlands) 21:100205; doi:10.1016/j.comtox.2021.100205.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58:267–288; doi:10.1111/j.2517-6161.1996.tb02080.x.

Tollefsen KE, Scholz S, Cronin MT, Edwards SW, Knecht J de, Crofton K, et al. 2014. Applying adverse outcome pathways (AOPs) to support integrated approaches to testing and assessment (IATA). Regulatory Toxicology and Pharmacology 70:629–640; doi:10.1016/j.yrtph.2014.09.009.

Villeneuve DL, Angrish MM, Fortin MC, Katsiadaki I, Leonard M, Margiotta-Casaluci L, et al. 2018. Adverse outcome pathway networks II: Network analytics. Environmental Toxicology and Chemistry 37:1734–1748; doi:10.1002/etc.4124.

Villeneuve DL, Crump D, Garcia-Reyero N, Hecker M, Hutchinson TH, LaLone CA, et al. 2014a. Adverse outcome pathway (AOP) development i: Strategies and principles. Toxicological Sciences 142:312–320; doi:10.1093/toxsci/kfu199.

Villeneuve DL, Crump D, Garcia-Reyero N, Hecker M, Hutchinson TH, LaLone CA, et al. 2014b. Adverse outcome pathway development II: Best practices. Toxicological Sciences 142:321–330; doi:10.1093/toxsci/kfu200.

Villeneuve D, Meek B, Viviani B, Burgdorf T, LaLone C, O'Brien J, et al. 2024. AOP developers' handbook version 2.7.

Young C. 2019. The difference between causal analysis and predictive models: Response to "comment on young and holsteen (2017)." Sociological Methods & Research 48:431–447; doi:10.1177/0049124118782542.

Zou T, Caton AJ, Koretzky GA, Kambayashi T. 2010. Dendritic cells induce regulatory t cell proliferation through antigen-dependent and -independent interactions. Journal of immunology 185:2790–2799; doi:10.4049/jimmunol.0903740.