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>

ABSTRACT

The difference between predictive and causal relationships has been a question in mechanistic toxicology. The adverse outcome pathways (AOPs) methodology is a mechanistic framework that supports the transition to non-animal methods in risk assessment. AOPs include molecular initiating events, key events, and adverse outcomes, linked by key event relationships (KERs) to infer higher-level outcomes from lower biological levels. To better understand the implications of causal and predictive KERs in AOP development, this study aimed to: 1) analyze the use of the terms describing KERs in key AOP guiding documents, and 2) examine their application in AOP development. Our text analysis on the two documents revealed that KER definitions could involve predictive, causal, or both. Examples were also found from studies of AOP development where KER definition is specified to be observational. The implications of these findings are discussed

in the context of KER evaluation. We conducted a literature analysis to assess how these terms are adapted across different stages of AOP development, noting that while the use of various KER definitions aligns with specific research objectives, the methodologies for establishing causality in some studies may be questionable.

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 non-animal methods for the prediction of adverse outcomes at higher biological levels (Villeneuve et al. 2014). 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. 2014). The utility of the extrapolation across events at different biological levels depends on the characterization of KERs (Villeneuve et al. 2023).

Causal inference is a cornerstone of science (Lipton and Ødegaard 2005). It 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). The gold standard of unbiased causal inference is randomized experiment (James et al. 2015) to approximate counterfactual estimates of treatment effects (Höfler 2005). When random experiments are

not feasible, inference must rely on observational data. The Bradford Hill criteria (Hill 1965) are recognized as an early framework for causal inference on observational data (Guzelian et al. 2005). However, this framework has limitations, including the reliance on authoritative judgments (Cox 2018). To overcome the biases from authoritative judgment in observational inference, evidence-based methodology employs the frameworks of weight of evidence (WoE) and systematic critical appraisal to evaluate best available evidence (Guzelian et al. 2005; James et al. 2015). The Modes of Action (MOAs) framework evaluates how well empirical evidence contributes to weight-of-evidence determinations of hypothesized mechanisms using an "evolved Bradford Hill considerations" (Boobis et al. 2006; Meek et al. 2014). MOA helps to infer relationships between biochemical events for better understanding of toxicological mechanisms. Traditional evidence-based toxicology primarily uses data from animal experiments (James et al. 2015). AOPs aim to make evidence-based evaluation on the KERs connecting MIE, KEs and AO. It has been suggested that the causality between stressors/ toxicants and the AO is necessary to the development of AOPs (Spînu et al. 2022).

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) and Quantative System Toxicology (Bloomingdale et al. 2017). The difference between causal and predictive models has been a source of debate and confusion (Young 2019). The term "causal" refers to the evaluation of a relationship under the framework of treatment effect evaluation (Young 2019). The term "predictive" term reflects the scope of

utility and application, and is not under the framework of treatment effect evaluation (Young 2019). The difference between "causal" and "predictive" terms can be 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 mentioned in the Bradford Hill criteria (Hill 1965). 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 appendix for these two points respectively.

Given the differences between the terms "causal" and "predictive" terms, the objectives of this article are to: 1) examine the use of terminology describing KERs across AOP documentations and 2) investigate the effects of adaptations of different terms on the development and application of AOPs.

2 METHODS

2.1 Terms for Key Event Relationships

We performed a text analysis on two AOP documentations including OECD's AOP Guidance (OECD 2017) and the AOP developer handbook (Villeneuve et al. 2023). The two

documents focus on setting the definitions, principles and guideline for AOP development. These two documents are selected for text analysis since their scopes align with the objectives of this study. Descriptions of KER are located with the following keywords: "key event relationship" OR "KER" OR ((relationship OR link) AND "key event"). We highlighted the examples identified from the sections where the definitions and principles of AOPs are set, since they set the context and references for the rest of the documents.

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 of terminology adaptation was organized by the stages of AOP and quantitative AOP (qAOP) development. Specifically, AOP development consists of identification of KEs, characterization of KERs and overall AOP evaluation, according to the AOP developer handbook (Villeneuve et al. 2023). qAOP development comprises stages including calibration of qAOP parameters (Zhou et al. 2024), prediction of the AO (Jeong et

al. 2018), derivation of point of departures (PODs) for the chosen KEs and the AO (Perkins et al. 2019a), and design of integrated testing strategies (Perkins et al. 2019a; Tollefsen et al. 2014).

3 RESULTS

3.1 Usage of terminology

The usage of "causal" and "predictive" varies within the two AOP documents (Table **1**), sometimes within the same page. For instance, KERs are described in two ways in the section to define KERs on page 24 of the handbook, where KERs could be "*the predictive relationships that facilitate extrapolation*" or "*a directed relationship (one-way arrow) linking an upstream "causing" event to a downstream "responding" event*". There are two instances where KERs are described to be both "causal" and "predictive". The concept of KERs is laid on page 6 of the handbook where a KER is "*a scientifically-based relationship that connects one key event to another, defines a causal and predictive relationship between the upstream and downstream event*". Since the usage of the two terms have different requirements, if both terms are to be used specifying KERs, the requirements of both need to be met.

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

		Terminology	
Location of KER description	Quote	Causal	Predictive
OECD guidance document			

		Terminology	
Location of KER description	Quote	Causal	Predictive
AOP definition	*a sequential chain of causally linked events at different levels of biological organisation*	√	
AOP handbook 2023 version			
Key term definition, 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*	√	√
Key term definition, KE, page 6 line 199	*Essentiality indicates that the KEs play a causal role in the pathway*	√	
AOP development tip #3, page 13 line 449	*an AOP is acceptable when there are multiple KEs, causally linked to the MIE and AO*	√	
KER description, page 24 line 901	*Within the AOP framework, the predictive relationships that facilitate extrapolation are represented by the KERs*		√
KER description, page 24 line 914	*a directed relationship (one-way arrow) linking an upstream "causing" event to a downstream "responding" event*	√	
KER evaluation protocol, page 29 line 1099	*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)*		√
KER evaluation protocol on biological plausibility, page 30 line 1134	*filter through much of the mechanistic detail to focus on what important causal events for the adverse outcome have predictive value for regulatory application*	√	√
KER evaluation	*assessment of changes in the upstream KE (KEupstream) leading to, or being associated with, a*		√

		Terminology	
Location of KER description	Quote	Causal	Predictive
protocol of empirical evidence, page 30 line 1140	*predictable subsequent change in the downstream KE (KEdownstream).*		
Overall AOP assessment of all KEs, page 38 line 1443	*in a manner consistent with its causal, and essential, role in the pathway*	√	
Overall AOP assessment of each KERs, page 40 line 1538	*Empirical support entails consideration of experimental data in terms of the associations between KEs*		√

From Table **1**, we see the term "causal" is employed to describe KERs derived from observational analyses. For instance, the example from Section 3 of the AOP developer handbook (Villeneuve et al. 2023) is a WoE protocol evaluating biological plausibility, empirical supporting evidence and current quantitative understandings of the KER. This protocol describes a KER of high WoE as "causal" when the KER is 1) biologically plausible and 2) backed by considerable empirical, quantitative evidence. However, the evaluation of biological plausibility is observational rather than experimental (Dailey et al. 2018). Therefore, when empirical evidence is not experimental, the evaluation under this protocol is observational. This means that without acknowledging these conditions, the "causal" term could be used under this protocol to describe KERs that are not necessarily causal.

There are challenges to obtain quantitative evidence needed for causal evaluation of KERs (Spînu et al. 2022). KERs evaluated on experimental data are not necessarily causal, due to

the inability to randomize upstream KEs, which needs to be understood under the context of evaluating multiple KEs.

Role of randomization. The key difference between experimental and observational analysis is the randomization of treatment status (Höfler 2005). Randomized controlled trials (RCTs) are considered as the gold standard in causal inference because it approximates unbiased estimates of average treatment effect (ATEs) through the randomization of treatment (Höfler 2005; Pearl 2009). The mechanism of RCTs is revisited in detail in the appendix.

Evaluation of multiple KEs. KER evaluation involves the characterization of relationships between multiple events, or multiple measures of the same event. To understand the mechanism of harmful effects of a toxicant, toxicologists evaluate the treatment effects on more than one biochemical events. The study of toxicological mechanisms, such as AOPs and MOAs (Boobis et al. 2006; Meek et al. 2014), characterizes the relationship across events at different biological levels. It is also not uncommon to measure the same biological event from multiple facets for better precision and cross validation, such as results from multiple assays or biomarkers. One goal of AOP's development is to understand the extent to which *in vitro* measurements could approximate *in vivo* measurements of the same biochemical event. For example, in order to distinguish between pharmacological responses at pharmacological doses and species-specific hepatotoxic responses at higher doses *in vivo* (Kienhuis et al. 2009), the effects of coumarin on gene expression in human hepatocytes is compared between *in vitro* to *in vivo* and *in vitro* rat studies.

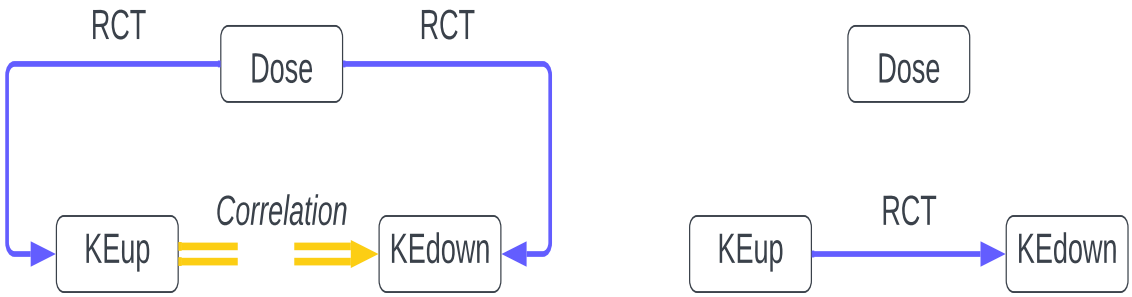

Figure 1: *Difference between observational and causal analysis of a key event relationship. Left panel is observational; right panel is causal.*

Consider a simple study design to evaluate the effect of a toxicant on two KEs (Figure **1**). The upstream KE is denoted as KE_{up}, the downstream one is denoted as KE_{down}. The effect of administering a given dose of the toxicant (termed the treatment) is evaluated, shown as the box named "dose". Two pairs of treatment-control data can be observed through this design for KE_{up} and KE_{down} respectively. The evaluation of treatment effects on KE_{up} and KE_{down} respectively is based on the random assignment of the treatment, as shown in the left panel of Figure **1**. Therefore, the average difference between the control and treatment groups provides an unbiased estimate of the treatment effect on KE_{up} and KE_{down} respectively.

Challenge to randomize KEs. A causal analysis of KER requires random assignments of KE_{up} instead of the treatment (right panel of Figure **1**) to obtain an unbiased ATE of KE_{up} on KE_{down}, However, KE_{up} itself is the consequence of the treatment. The onset of KE_{up} is not randomly assigned among subjects. Direct evaluation of 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. For example, iis possible that there is no causation between KE_{up} and KE_{down}. The observed changes in KE_{up} and KE_{down} is attributable to a third, unobserved event. The estimated effect of KE_{up} on KE_{down} would be biased without isolating the effect of confounding. Therefore, despite the involvement of experimental data, the average difference in KE_{down} conditioning on the status of KE_{up} could be biased. It faces the same problem to infer or counter the reverse-causality of KE_{down} on KE_{up}.

In summary, there can be causal as well as observational inference of KERs drawn from data collected through experiments. It is important to acknowledge that results from observational analysis are statistical associations rather than causal relationships.

3.2 Adaptation of terminology in AOP development and application

We list examples and suggestions of adapting the term "causal", "predictive" and "observational" for KER definition for different purposes and tasks in AOP development and application (Table **2**). We include the term "observational" because we found examples explicitly (Velthuis et al. 2023) or implicitly (MacKay et al. 2013) acknowledging their KER definition as observational. There were no examples of the adaptation of the term "observational" in describing KERs for the prediction of AO, derivation of PODs and design of testing strategies.

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

Purpose		Given the purpose, one example of adapting term ___ for KER definition is to:		
Application Domain	Task	Causal	Predictive	Observational
AOP development	Identify KE	Use quasi-experimental methods to approximate unbiasedness [1]	Consider event of good predictive performance as KE [4]	*Use modified Hill criteria*
	Characterize KER		Consider edge in a network model as KER [4]	Apply KER WoE evaluation protocol in AOP developer handbook [9]
	Evaluate AOP	Use DAGs to express established causality [2]		Apply AOP WoE evaluation in AOP developer handbook [9]
qAOP development and application	Calibrate qAOP	Consider KERs in developed AOP as causal [3]	*Calibrate parameters to optimize performance metrics [6]*	Evaluate robustness of parameterization by consistency to Hill Criteria [10]
	Predict AO		Predict AO based on strong predictive KE and KER [7]	
	Derive POD		Build predictive response-response model [5]	
	Design testing strategy		Use conditional dependency or probability to determine event for testing [8]	

Suggestions in italic

KE: key events; KER: key event relationship; DAG: directed acyclic graph; AO: adverse outcome; AOP: adverse outcome pathway; POD: point of departure; WoE: weight of evidence;

Purpose		Given the purpose, one example of adapting term ___ for KER definition is to:		
Application Domain	Task	Causal	Predictive	Observational

[1] Landfors et al. 2023, [2] Spinu et al. 2022, [3] Burgoon et al. 2020, [4] Villeneuve et al. 2018, [5] Perkins et al. 2019, [6] Zhou et al. 2024, [7] Jeong et al. 2018, [8] Tollefsen et al. 2014, [9] Villeneuve et al. 2023, [10] Song et al. 2023

3.2.1 Adaptation of "causal" term

By the completion of this paper, there has been no practical solution to randomize KE_{up} for the characterization of KERs. Quasi-experimental methods approximate unbiased ATE on observational data, conditional on the satisfaction of specific assumptions (Angrist and Pischke 2009). These methods have been employed in various fields such as economy, social sciences, psychology and epidemiology (Angrist and Pischke 2009; Bagiella et al. 2015; Fréchette et al. 2022). Targeted mutation with the instrumental variable method has been implemented in the identification of KEs and characterization of KERs (Landfors et al. 2023). Known causality can be propagated across an AOP using Directed Acyclic Graphs (DAGs) (Spînu et al. 2022).

When it comes to the development and application of qAOPs, KERs are often taken for granted as causal (Burgoon et al. 2020) in spite of the potential biases in the ATEs. The AOP developer handbook defines KERs with "high weight of evidence" as causal (Villeneuve et al. 2023). This causality is further integrated into overall AOP evaluation to extrapolate the relationship between MIE and AO.

3.2.2 Adaptation of "predictive" term

Predictive models have been widely applied in AOP development, including machine learning (Gou et al. 2023) and 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. 2019a). Similar methods have been applied to screen for genes and genomic modules associated with significant differential gene expressions (Callegaro et al. 2021). Graphical network models represent biochemical events including the AO as nodes and their links as connecting edges (Villeneuve et al. 2018). The 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 consist of deterministic and probabilistic nodes, and the undirected, directed, acyclic edges, that have been involved in existing studies of AOP development and application (OECD 2017; Perkins et al. 2019a). The identified KEs and KERs could be parameterized (Zhou et al. 2024) to further optimize the model's predictive performance given an objective, including the prediction of AO, derivation of PODs and development of testing strategies.

3.2.3 Adaptation of "observational" term

Table 3: *Proposed adaptation of Hill's criteria for evaluating observational KERs based on current practice*

	Original Hill's criteria	Modified Hill's criteria
Scope of inference	Relationship between disease and outcome	Relationship between two KEs
Strength	*A small association does not mean that there is not a causal effect, though the larger the association, the more likely that it is causal.*	A high association of a KER does not increase the likelihood of causality between Kes, although a higher association with bias adjustment can indicate less bias in the evidence for the KER.

	Original Hill's criteria	Modified Hill's criteria
Scope of inference	Relationship between disease and outcome	Relationship between two KEs
Consistency	*Consistent findings observed by different persons in different places with different samples strengthens the likelihood of an effect.*	Same.
Specificity	*Causation is likely if there is a very specific population at a specific site and disease with no other likely explanation. The more specific an association between a factor and an effect is, the bigger the probability of a causal relationship.*	If no biologically plausible alternative explanation could be raised, the KER estimation is more likely to be causal.
Temporality	*The effect has to occur after the cause (and if there is an expected delay between the cause and expected effect, then the effect must occur after that delay).*	The upstream event has to occur before the downstream event.
Biological gradient	*Greater exposure should generally lead to greater incidence of the effect. However, in some cases, the mere presence of the factor can trigger the effect. In other cases, an inverse proportion is observed: greater exposure leads to lower incidence.*	The dose threshold of the upstream event should be no greater than that of the downstream event.
Plausibility	*A plausible mechanism between cause and effect is helpful (but Hill noted that knowledge of the mechanism is limited by current knowledge).*	Same.
Coherence	*Coherence between epidemiological and laboratory findings increases the likelihood of an effect. However, Hill noted that "lack of such [laboratory] evidence cannot nullify the epidemiological effect on associations".*	Coherence between animal and non-animal evidence increases the likelihood of causality and strengthens external validity.
Experiment	*Occasionally it is possible to appeal to experimental evidence.*	Randomization of upstream KE is practically difficult, but is possible to explore quasi-experimental evidence.
Analogy	*The use of analogies or similarities between the observed association and any other associations.*	The use of analogy or similarity between the observed KER and endorsed or biologically plausible KERs.

	Original Hill's criteria	Modified Hill's criteria
Scope of inference	Relationship between disease and outcome	Relationship between two KEs
The more criteria that are met...	…the higher the likelihood for the effect to be causal.	…the lower risk of bias in the evidence for the KER.

Observational analysis employs non-experimental methods to mitigate biases in ATEs (Angrist and Pischke 2009). The Bradford Hill criteria (Hill 1965) was first developed for epidemiological inference between a presumed cause and an observed public health effect (Hill 1965), including strength, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment and analogy. We propose a modified version of Hill criteria (Table **3**) tailored for the context of KER evaluation based on current practice (Becker et al. 2015; Perkins et al. 2019b). A modified strength criterion is proposed similar to a previous statement (Spînu et al. 2022) that a high correlation between two KEs does not necessarily increase the likelihood of causality, as data and statistics alone cannot answer causal questions (Pearl and Mackenzie 2018). The biological gradient criterion defines that the threshold dose of KE_{up} is expected to be no greater than that of KE_{down}.

The KER evaluation protocol in the AOP developer handbook (Villeneuve et al. 2023) includes criteria for evaluating empirical evidence similar to our considerations on the criteria of temporality, biological gradient and plausibility. The biological gradient criterion is expanded in the handbook to the evaluation of dose concordance and incidence concordance. The KER evaluation protocol has been employed to evaluate the robustness of parameterization in the calibration of qAOPs. The temporal and dose concordances are checked to evaluate the quality of fitted parameters and estimated qAOP (Song et al. 2023). There are two major differences between our modified Hill criteria and the handbook's

KER evaluation protocol. Firstly, our modified Hill criteria acknowledges that the results are not necessarily causal and there could be potential biases even if all criteria are met. The handbook's protocol concludes that a KER has high WoE and is causal when 1) has biological plausibility 2) the supporting evidence passes the concordance checks 3) is supported by quantitative evidence. With our modified criteria, we propose that a KER must 1) has biological plausibility and 2) pass the temporality criteria to be considered, and has a relatively higher WoE and lower risk of bias (Sterne et al. 2019) if it passes any other criterion. The reason to not require biological gradient is that there is ambiguity in the choice of thresholds for the two KEs. Secondly, the handbook's protocol allows the aggregation of results to the overall AOP WoE evaluation. Our modified Hill criteria have not been extended to include the aggregation of causality across an AOP.

4 DISCUSSIONS

4.1 Mismatch between methodology and terminology

Based on our results on how the terminology has been adapted in the development of AOPs and qAOPs, there could be three types of mismatch between the terminology and the method of adaptation (Table **4**).

Table 4: *Three types of mismatches between KER terminology and their adaptation in AOP development*

Methods adapted in AOP development	Term	
	Predictive	Causal
Predictive		2

Methods adapted in AOP development	Term	
	Predictive	Causal
Causal	1	
Observational	1	3

Type 1 and 2 assume the mismatch between predictive models and treatment effect framework. The goal of the treatment effect framework is to approximate unbiasedness (Young 2019) using experimental methods for causal analysis and non-experimental/quasi-experimental methods for observational analysis (Angrist and Pischke 2009). The objective of predictive models is to improve predictive performance. Type 1 specifies a situation when the term "predictive" is used on results derived from the treatment effect framework, including causal and observational inference. Type 2 is the opposite when the term "causal" is employed on the outputs of predictive models.

To elaborate, the practical difference between predictive models and the treatment effect framework is discussed. Consider a linear regression with multiple covariates, whose parameters are optimized using maximum likelihood estimation (MLE). 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, such as R square values and Akaike Information Criteria (AICs) (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 MLE) (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 R square and AICs (Brookhart et al. 2010; Davies et al. 2013). Including predictors of local significance may not improve global R square or AICs (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).

Type 1 mismatch was not found in our literature analysis. To meet the need of AO prediction (OECD 2017), predictive analysis is performed on KERs with established causality. Bayesian approaches have been suggested to integrate known causality of KERs and prior knowledge on KEs from experts into predictive models (Spînu et al. 2022). Furthermore, known causation between predictors and the outcome could be incorporated in the predictive model by adjusting the loss/utility function, such as Lasso regularization (Tibshirani 1996). Since these KERs are both causal and predictive, it is possible that emphasizing their predictive utility better aligns with the objectives of specific research.

Type 2 mismatch could be demonstrated through possible misinterpretations of probabilistic network models as causal. Probabilistic network models (Koller and Friedman 2009) are based on 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 ATEs for causal inference. In Bayesian modelling, DAGs 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. BNs 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 BN (Kaikkonen et al. 2021; Spinu et al. 2020) comes misconceptions of to what extent a BN are causal, which might come from the use of a DAG. There are studies where results of BN were interpreted as causal without employing quasi-experimental bias reduction (Chen and Pollino 2012; Jeong et al. 2018; Zou et al. 2010). For evaluating the uncertainty in a BN, beyond the inherent uncertainty embedded within the probability model, it is crucial to consider the process and context in which the BN is applied (Sahlin et al. 2021).

Type 3 mismatch exists when causality is claimed on the KERs from observational analysis where bias has not been properly accounted for. The AOP developer handbook's KER evaluation protocol and AOP evaluation protocol (Villeneuve et al. 2023) adapt the term "causal" on results derived from observational analysis. Therefore, these usages of the term "causal" in the handbook's protocol could be type 3 mismatches if the biases are not appropriately adjusted, which impacts the validity of the compliance review for OECD's AOP endorsement process (OECD 2021). To avoid type 3 mismatch, it is vital to recognize that KER analysis might be observational instead of causal. Existing KER studies need to be ascertained whether they were causal or observational. For the KERs derived from observational studies, their results should not be entirely discredited. Instead, a new

direction of discussion opens that KERs from observational studies should be evaluated for 1) the robustness (Frank et al. 2013) and sensitivity (Oster 2019; Rosenbaum 2005) of results and 2) whether the disclosed information allows for retrospective quasi-experimental bias adjustment. More research is needed to formulate a systematic protocol to decide what is needed to enable quasi-experimental analysis. Future studies on AOP development and KER characterization have the opportunity to achieve causal conclusions by exploring the possibility to randomize upstream events and to adjust for omitted variable biases during the designing stage of experiments.

KER analysis employing targeted mutation methods has much less bias (see SI), but could still be observational due to omitted variable bias. The links between genes and proteins by genome-wide association studies are associations 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).

4.2 Definition of "causal"

The AOP developer handbook defines the essentiality and necessity of KEs as one of the key aspects of KERs that "*essentiality indicates that the KEs plays a causal role in the pathway*"

(Villeneuve et al. 2023). For instance, Table 1 on page 6 of the handbook emphasizes the importance of KERs to AOPs as "*if a given KE is prevented or failed to occur, progression to subsequent KEs in the pathway will not occur*". We argue that defining the term "causal" as essential/ necessary could be practically difficult and conceptually questionable.

Firstly, necessity and essentiality are specialized terms to describe the conditional relationship between two statements. To claim the 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 essentiality and necessity requires a falsity test. Any case where the occurrence of B without the occurrence of A falsifies the necessity statement. 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.

Therefore under the context of AOPs, essentiality indicates that KE_{down} cannot occur at all if KE_{up} has not occurred. It is practically challenging to define causality on the basis of essentiality. The validation of essentiality of KE_{up} for KE_{down} requires an exhaustive testing of all possible scenarios where KE_{down} occurs for the confirmation of the involvement of KE_{up} in all the scenarios. Moreover, the practicability of the essentiality definition is further challenged by an example in the AOP developer handbook that could counter the definition of essentiality. At page 13 of "*development tips 3*", a branching AOP is presented as "*multiple KEs, causally linked to the MIE and AO that are occurring concurrently and acting in concert to drive the downstream effects*". Under the essentiality definition, all the branched KEs must occur simultaneously to enable the occurrence of AO. Obviously, such a statement is vulnerable to observations of inactive upstream KEs.

Specifying the definition of causal on the basis of necessity could be conceptually questionable under the context of AOP. The definition of essentiality and necessity, along with the quote from the handbook above, are inconsistent with the principles for AOP development presented in the handbook. The handbook distinguishes AOPs from MOAs (Boobis et al. 2006) that AOP presents only a selection of "check-point" events, instead of a comprehensive description of every biological activity involved. The "check-point" principle is to select events at each biological level with the largest magnitude of impact on the perturbation of its subsequent KEs or AO (Villeneuve et al. 2023). This principle acknowledges the potential existence of factors mediating the downstream events at a marginal magnitude without the check point. Because AOPs only include check-point events as KEs, this principle is against the essentiality and necessity definition.

Moreover, there are AOP examples in the AOP handbook where multiple AOPs lead to the same AO. Under the essentiality definition, these AOPs leading to the same AO should either merge into one gargantuan AOP, or compete against each other to determine which one is truly essential to the perturbation of the AO. The conceptual conflict becomes worse when these AOPs with shared AO also have partially shared KEs. An example of *"parallel AOPs supporting the same AO"* cited in the AOP handbook is the development of AOPs leading to impaired egg production in fish (Ankley et al. 2010). This example is demonstrative for AOP development but can also be considered to counter the definition of essentiality. The KE of reduced vitellogenin production is shared by three AOPs leading to impaired ovulation in female fish in Figure 4 of the original article: 1) antagonism of estrogen receptor 2) inhibition of of 17β-estradiol synthesis and 3) indirect reduction of 17β-estradiol production by feedback inhibition of steroidogenesis. The preceding KEs to

reduced vitellogenin production differ across the three AOPs as well. Therefore, the essentiality definition would question whether one or all preceding KEs is truly essential to reduced vitellogenin production and whether one or all the AOPs is truly essential to the AO of impaired ovulation in fish.

5 CONCLUSIONS

KERs are defined to be predictive, causal, or both, as found by our text analysis of two AOP guiding documents. There are also examples from the studies of AOP development that KER definition is acknowledged to be observational explicitly or implicitly. We demonstrated the difference among predictive, causal, both predictive and causal and observational types of relationships under the context of KER evaluation. Adaptation of these terms in KER definition varies across the stages of AOP development. The term predictive has been widely employed in AOP development to support predictive models such as machine learning and graphical networks. However, it requires caution to interpret findings from predictive models, especially when the research question addresses the causality in KERs. Employing the causal term for KER evaluation faces the challenge to perform randomization on observed outcomes, under which scenario KER evaluation is observational. The exploration of quasi-experimental methods in AOP development is currently limited. Claiming causality on the results from observational analysis could underestimate the potential risk of bias. Finally, it is questionable to define causality on essentiality under the context of KER evaluation. Our investigation on the usage and adaptation of terms describing KERs clarifies the confusion around these terms and supports the development and application of AOPs.

6 REFERENCE

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

Ankley GT, Bennett RS, Erickson RJ, Hoff DJ, Hornung MW, Johnson RD, et al. 2010. Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. Environmental Toxicology and Chemistry 29:730–741; doi:https://doi.org/10.1002/etc.34.

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.

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.

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.

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.

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.

Dailey J, Rosman L, Silbergeld EK. 2018. Evaluating biological plausibility in supporting evidence for action through systematic reviews in public health. Public Health 165:48–57; doi:10.1016/j.puhe.2018.08.015.

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.

Frank KA, Maroulis SJ, Duong MQ, Kelcey BM. 2013. What would it take to change an inference? Using rubin's causal model to interpret the robustness of causal inferences. Educational Evaluation and Policy Analysis 35:437–460; doi:10.3102/0162373713493129.

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.

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.

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

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

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.

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.

Kienhuis AS, Poll MCG van de, Dejong CHC, Gottschalk R, Herwijnen M van, Boorsma A, et al. 2009. A toxicogenomics-based parallelogram approach to evaluate the relevance of coumarin-induced responses in primary human hepatocytes in vitro for humans in vivo. Toxicology in Vitro 23:1163–1169; doi:10.1016/j.tiv.2009.06.005.

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.

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

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. 2021. *Guidance document for the scientific review of adverse outcome pathways*. Organisation for Economic Co-operation; Development.

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

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.

Oster E. 2019. Unobservable selection and coefficient stability: Theory and evidence. Journal of Business & Economic Statistics 37:187–204; doi:10.1080/07350015.2016.1227711.

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.

Pearl J, Mackenzie D. 2018. *The book of why: The new science of cause and effect*. Basic books.

Perkins EJ, Ashauer R, Burgoon L, Conolly R, Landesmann B, Mackay C, et al. 2019a. 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.

Perkins EJ, Gayen K, Shoemaker JE, Antczak P, Burgoon L, Falciani F, et al. 2019b. Chemical hazard prediction and hypothesis testing using quantitative adverse outcome pathways. ALTEX - Alternatives to animal experimentation 36:91–102; doi:10.14573/altex.1808241.

PrecisionTox. 2024. Publications-PrecisionTox.

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

Rosenbaum PR. 2005. Sensitivity analysis in observational studies.

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.

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.

Sterne JA, Hernán MA, McAleenan A, Reeves BC, Higgins JP. 2019. Assessing risk of bias in a non-randomized study. Cochrane handbook for systematic reviews of interventions 621–641.

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.

Velthuis K, Poppelaars F, Ten Klooster PM, Vonkeman HE, Jessurun NT. 2023. Impact of adverse drug reactions on the treatment pathways of early rheumatoid arthritis patients: A prospective observational cohort study. Expert opinion on drug safety 22:753–762; doi:10.1080/14740338.2023.2194628.

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. 2014. Adverse outcome pathway (AOP) development i: Strategies and principles. Toxicological Sciences 142:312–320; doi:10.1093/toxsci/kfu199.

Villeneuve D, Meek B, Viviani B, Burgdorf T, LaLone C, O'Brien J, et al. 2023. AOP developers' handbook.

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.

Zhou Z, Sahlin U, Cronin M, Madden J, Firman J, Zgheib E. 2024. Methodologies to calibrate a qAOP.

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.