CAUSAL, PREDICTIVE OR OBSERVATIONAL? DIFFERENT UNDERSTANDINGS OF KEY EVENT RELATIONSHIPS FOR ADVERSE OUTCOME PATHWAYS AND THEIR IMPLICATIONS ON PRACTICE

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>

10 # 1 INTRODUCTION

11 ## 1.1 Background

12 Regulatory toxicology and environmental risk assessment in the twenty-first century need
13 to handle the assessment of ever-increasing numbers of chemicals, and at the same time
14 adhere to the 3R principles to reduce, refine and replace animal experiments (Pallocca et al.
15 2022). The Adverse Outcome Pathways (AOPs) methodology facilitates the extrapolation of
16 evidence at lower biological levels from animal and non-animal methods for the inference
17 of adverse outcomes at higher biological levels (Villeneuve et al. 2014a). An AOP is defined
18 as an evolving conceptual framework to organize existing knowledge pertaining to a
19 molecular initiating event (MIE), an apical adverse outcome of interest (termed adverse
20 outcome; AO), key measurable/ observable events (termed key events; KEs), and the
21 linkages among MIE, AO and KEs (termed key event relationships; KERs) (Villeneuve et al.
22 2014a). A list of definitions and abbreviations is presented in the supplementary
23 information (SI). The robustness and reliability of the extrapolation across events at
24 different biological levels therefore depends on the characterization and inference of KERs
25 (Villeneuve et al. 2024), which could be causal, observational and predictive.

26 ## 1.2 Causal Inference and Its Challenges in Key Event Relationship Evaluation

27 Causal inference pursues the unbiased understanding of how one condition, the cause,
28 affects another condition, the outcome, free from the threats to internal validity (Guzelian
29 et al. 2005). Randomized experiments are the gold standard for causal inference (Guzelian
30 et al. 2005; James et al. 2015) as they can approximate unbiased counterfactual estimates
31 of treatment effects (Höfler 2005; Pearl 2009). Toxicological studies typically implement

randomized experiments in controlled laboratory conditions, which reveal unbiased estimates of the association between the randomized assignment (e.g. dose) to the compound of interest and the experimental outcome (e.g. response), thereby establishing a causal relationship(Höfler 2005). The mechanism of randomized controlled trials is reviewed in the SI.

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 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 upon acetaminophen overdose.

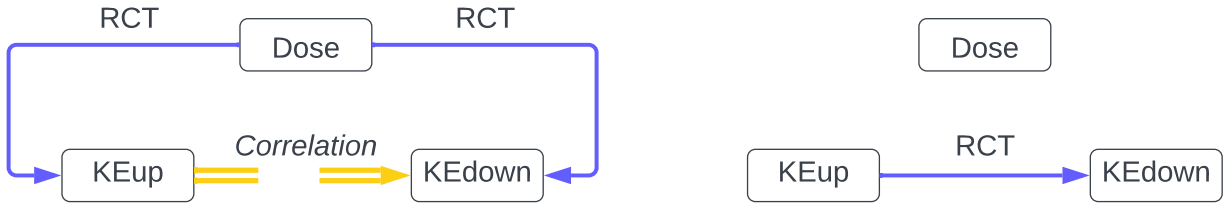

Figure 1: Directed acyclic graph of key event relationship evaluation. RCT: randomized controlled trials. Left shows that treatment effects are causal on both key events respectively; right shows what is needed for a causal evaluation

1.3 Distinguishing Causal and Predictive Inference under the Context of Adverse Outcome Pathways

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 quantitative 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 (Angrist and Pischke 2009), evidence-based frameworks). In contrast, the objective of predictive inference is to make accurate predictions. The term "predictive" 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 (Russo and Williamson 2007), whereas the opposite is the error of reverse-causality. 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 of 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.

119 Furthermore, practical differences between predictive and the treatment effect
120 frameworks exist in many commonly used methods. Consider a linear regression with
121 multiple covariates, whose parameters are optimized using maximum likelihood
122 estimation. This regression can be used for both prediction and causal inference (Young
123 2019). When this regression is used as a predictive model, the goal is to improve the
124 model's predictive performance (Young 2019), through the selection and optimization of
125 the design matrix. When this regression is used as a causal model, bias adjustment is
126 performed on the data prior to model construction, such as control selection (Cinelli et al.
127 2022) and backdoor elimination (Pearl 2009). The goal is to use the adjusted data and
128 identify predictors that have significant coefficients (p values using maximum likelihood
129 estimation and null hypothesis testing) (Angrist and Pischke 2009). The same model works
130 differently for prediction and treatment effect evaluation. On the one hand, bias adjustment
131 on data may not improve predictive performance (Brookhart et al. 2010; Davies et al.
132 2013). Including predictors of local significance may not improve global performance
133 (Obermeyer and Emanuel 2016). On the other hand, optimizing the design matrix can not
134 reduce the bias without prior bias adjustment (Chen et al. 2021).

135 The distinction between causal, observational and predictive inference is particularly
136 relevant to context of toxicology and AOP development. This difference should be
137 considered in the assembly of KEs and KERs. Traditionally, KEs and KERs are characterized
138 among measured/ observed events from existing data, which is retrospective and
139 evidence-based. In contrast, recent *in silico* predictive models enable prospective forecast
140 of KEs and KERs for new compounds with minimal structural analysis (Hemmerich and
141 Ecker 2020). The distinction between causal, observational and predictive inference should

142 also be considered taking the difference between MOAs and AOPs. A major distinction
143 between MOAs and AOPs is that MOAs are chemical specific while AOPs are not (Villeneuve
144 et al. 2014a; Villeneuve et al. 2024). Evaluation of KERs under the MOA framework (Boobis
145 et al. 2006) involves the evolved MOA Hill considerations (Meek et al. 2014), where the
146 confidence in KER causality is crucial along the MOA of the compound of interest. In the
147 context of AOP development, it has been suggested that the causality is necessary between
148 the AO and compounds on all AOP models including predictive ones (Spînu et al. 2022).
149 However, it remains unclear whether it is necessary to require causality in all KERs in a
150 chemical agnostic AOP. Given the chemical agnostic principle of AOPs, the supporting data
151 for AOP development may include a wide range of compounds, making it common to
152 observe varying KE activities across compounds. Therefore, it is crucial to consider the
153 appropriate application of causal, observational and predictive inference in AOP
154 development.

155 **1.4 Study Objectives**

156 Given the difference between the terms "causal", "observational" and "predictive" and its
157 relevance to KER characterization, we hypothesize that there could be misuse of the terms
158 and it hinders the effective evaluation of KERs. To test this hypothesis, the objectives of this
159 study are to: 1) examine the use of terminology describing KERs across AOP
160 documentations and 2) investigate the adaptation of different terms in the development
161 and application of AOPs. Achieving these objectives may facilitate a continuous
162 development of AOP guidelines and a pragmatic implementation of AOP framework in
163 academic, regulatory and industrial contexts.

164 **2 METHODS**

165 **2.1 Terms for Key Event Relationships**

166 We performed a text analysis on two AOP documentations focused on setting the
167 definitions, principles and guidelines for AOP development: AOP developer's handbook
168 (Villeneuve et al. 2024) (abbreviated as the handbook) and OECD Guidance Document for
169 the Use of Adverse Outcome Pathways in Developing Integrated Approaches to Testing and
170 Assessment (OECD 2017) (abbreviated as OECD guidance thereafter). For a comprehensive
171 analysis on the descriptions of KERs, manual searches were performed in the handbook
172 and OECD guidance with the following keywords: "key event relationship" OR "KER" OR
173 ((relationship OR link) AND "key event"). We also noted the sentences before and after the
174 located keywords for context.

175 **2.2 Adaptation of terminology**

176 We searched journal articles published by May 2024 to identify how different terminology
177 has been adapted in development and application of AOPs. Our search was performed
178 using online scholarly databases such as Google Scholar and PubMed. We gave preferences
179 to sources from the "Animal-free Safety assessment of chemicals: Project cluster for
180 Implementation of novel Strategies" (ASPIS) research cluster, including: ONTOX (ONTOX
181 2024), RISK-HUNT3R (RISK-HUNT3R 2024) and PrecisionTox (PrecisionTox 2024). We
182 prioritized information from ASPIS due to the alignment between this article's emphasis on
183 AOPs and ASPIS's perspective on non-animal methodologies. These searches aimed to
184 identify exemplary cases to illustrate typical scenarios where the terms describing KERs
185 have been adapted, rather than to conduct a systematic review of all relevant literature.

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

194 **3 RESULTS**

195 **3.1 Usage of terminology**

196 KER descriptions in the OECD guidance are primarily based on the AOP developer's
197 handbook, with minor editorial changes for the use in integrated testing strategies.
198 Therefore, results of our text analysis focus on KER descriptions identified from the
199 handbook, highlighting the text examples located from the sections where the definitions
200 and principles of AOP development are set. Results show that with a clear distinction
201 between the terms "causal" and "predictive", the handbook's usage of them varies across
202 sections (Table 1). For instance, in the handbook Section 3 KER Description on page 25,
203 KERs are described in different ways: as predictive *"Within the AOP framework, the*
204 *predictive relationships that facilitate extrapolation are represented by the KERs"* or as
205 causal *"KERs are always described in the form of a directed relationship (one-way arrow)*
206 *linking an upstream "causing" event to a downstream "responding" event"*. In some instances

207 KERs are described with both "causal" and "predictive", e.g., on page 6 Table 1 Definitions
208 of Key Terms that *"A scientifically-based relationship that connects one key event to another,*
209 *defines a causal and predictive relationship between the upstream and downstream event"*.
210 Compared to instances where "causal" or "predictive" are used separately, the inclusion of
211 both terms suggests the consideration of dual requirements in KER description. The dual
212 requirements do not necessarily imply a conflict against separate use of terms. When both
213 causal and predictive inference are needed for KER description, one or both may be
214 emphasized at different stages of AOP development.

215 The dual requirements are also evident in the handbook's protocol for evaluating KER
216 supporting evidence. The handbook evaluates KER supporting evidence (Section 3G)
217 through 1) biological plausibility, 2) empirical evidence and 3) current quantitative
218 understandings. The handbook's guidance to assess biological plausibility (annex 1) is
219 adapted from the evolved MOA Hill considerations (Meek et al. 2014), which is
220 observational as mentioned in the Introduction of this paper. Empirical evidence is
221 assessed (handbook page 31 subsection ii) for predictive utility that *"assessment of changes*
222 *in the upstream KE (KEupstream) leading to, or being associated with, a predictable*
223 *subsequent change in the downstream KE (KEdownstream)"*. Moreover, within the same
224 section, the uncertainty and inconsistency of empirical evidence is evaluated that *"while*
225 *there are expected patterns of concordance that support a causal linkage between the KEs in*
226 *the pair, it is also helpful to identify experimental details that may explain apparent*
227 *deviations from the expected patterns of concordance"*. The handbook's evaluation of current
228 quantitative understandings of KERs, considering the weights of empiricial evidence
229 supporting KER causality, is dedicated to prediction *"to define the precision with which the*

230 *state of the downstream KE can be predicted from knowledge of the state of the upstream KE"*

231 (page 34). Therefore, it seems that the handbook sets the definitions, descriptions and

232 protocols for evaluation of KERs based on a dual focus of causal and predictive inference.

233 The term "observational" or the more commonly recognized "causal estimates" does not

234 appear in the handbook, nor the OECD guidance.

235 *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	*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)*		√

		Terminology	
Location of KER description	Quote	Causal	Predictive
Evidence, page 29 line 1126			
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, Uncertainties and Inconsistencies, page 32 line 1242	*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.*	√	
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.*		√

236 **3.2 Adaptation of terminology in AOP development and qAOP application**

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

238 *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 between MIE and AO		As a result of known gaps and uncertainties (Bal-Price et al. 2015; Becker et al. 2015)	
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	Same as in observational but

Stage	Characteristics	Causal	Observational	Predictive
			magnitude and/or duration of change in upstream KE needed to evoke change in downstream KE (Becker et al. 2015)	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)
	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

239 The handbook's dual requirement on both causal and predictive KERs provides a

240 comprehensive framework for AOP development. Examples identified from published

241 literature are categorized by which term among "causal", "observational" and "predictive"

242 was adapted, and displayed throughout stages (Villeneuve et al. 2014a) of putative,
243 qualitative and quantitative AOP development (Table 2).

244 Adapting the term "causal" in the strict sense would require experiments implementing
245 direct randomization of upstream KEs. However, no examples of practical implementation
246 could be identified in the literature. The alternative approach in published studies is
247 targeted blocking experiments to infer the impact of KE_{up} on KE_{down} that a reduction or
248 complete absence of a KE_{down} when KE_{up} is blocked (Meek et al. 2014). Empirical evidence
249 from targeted blocking experiments are direct support and contribute to a high level of
250 confidence for evaluating the essentiality of KEs (Meek et al. 2014; Villeneuve et al. 2014b;
251 Villeneuve et al. 2024). The development and application of qAOPs often assumes that KEs
252 are essential (Jensen et al. 2023) and KERs are causal (Burgoon et al. 2020; Gou et al.
253 2023).

254 Observational analyses of KEs, KERs and overall AOP could involve weight-of-evidence
255 (WoE) assessments of supporting evidence and quasi-experimental assessments. The
256 evolved MOA Hill considerations (Meek et al. 2014) are adapted in the handbook and the
257 OECD guidance for the WoE assessment of the essentiality of KEs , the biological
258 plausibility of KERs, the concordance of KER empirical evidence and the overall AOP level
259 of confidence (OECD 2017; Villeneuve et al. 2024). Best practice has been recommended
260 for the adapted WoE assessments (Villeneuve et al. 2014b) to categorize confidence levels
261 as high, moderate or low (Becker et al. 2015). Characterizing known gaps and uncertainties
262 that modulate incomplete linkages between MIE and AO could further enhance the weights
263 of evidence in the overall AOP (Bal-Price et al. 2015; Villeneuve et al. 2024). When high

264 WoE is achieved, qualitative and quantitative descriptions of KE measurement methods
265 and their accuracy and precision are crucial (Villeneuve et al. 2014b). These descriptions
266 support the characterization and calibration of predictive modeling of relationships among
267 the sequence of KEs in the overall AOP (Becker et al. 2015). In addition to WoE
268 assessments, the instrumental variable method, a type of quasi-experimental design, shows
269 promise in providing strong evidence supporting KER causality (Landfors et al. 2023).

270 Predictive models are primarily involved in the putative and quantitative stages (OECD
271 2017). Assembly of KEs and KERs in the putative stage can be derived using machine
272 learning (Gou et al. 2023; Hemmerich and Ecker 2020) or probabilistic graphical network
273 models (Villeneuve et al. 2018). KEs could be screened from a pool of candidate molecular
274 and cellular events using feature selection techniques (Perkins et al. 2019). Similar
275 methods have been applied to screen for genes and genomic modules associated with
276 significant differential gene expressions (Callegaro et al. 2021; Hemmerich and Ecker
277 2020). Topological metrics are evaluated to identify nodes and edges that best improve
278 overall predictive performance on the AO as KEs and KERs, respectively (Villeneuve et al.
279 2018). Graphical network models might consist of deterministic and probabilistic nodes,
280 and undirected, directed, acyclic edges, that have been involved in existing studies of AOP
281 development and application (OECD 2017; Perkins et al. 2019; Villeneuve et al. 2018).
282 Quantitative AOP development focuses on the derivation of response-response models
283 [Villeneuve2024]. Studies of qAOP typically assume the causality of KERs has been
284 established (Burgoon et al. 2020; Gou et al. 2023; Jensen et al. 2023), supported by the
285 essentiality of KEs, the biological plausibility of KERs and the concordance in KER empirical
286 evidence. Furthermore, to assure the reliability of qAOPs within relevant applicability

287 domains, the handbook includes a guidance (annex 2) to evaluate the WoE of qAOPs by not
288 only predictive performance, also considerations for known modulating factors, feedback
289 loops and generalizability across the applicability domains. Examples of possible
290 regulatory application of qAOPs include prediction of AO given MIE (Jeong et al. 2018),
291 derivation of omics (Song et al. 2023) and *in vitro* (Perkins et al. 2019) PODs, and
292 optimization of testing strategies (OECD 2017).

293 In summary, the handbook's nuanced usage of terms is reflected in AOP development that
294 focus of adaptation varies across stages of AOP development. Given the challenge of
295 obtaining experimental evidence for achieving causality, targeted block experiments offer a
296 practical approach to balance weights of evidence and feasibility. Best practice (Becker et
297 al. 2015; Villeneuve et al. 2014b) has been recommended to implement observational
298 inference using WoE assessments based on the evolved MOA Hill considerations (Meek et
299 al. 2014). The stage of putative AOP development has a dual focus on observational and
300 predictive evaluations. The qualitative stage focuses on the WoE of evidence supporting
301 KEs, KERs and the overall AOP. The stage of qAOP application and application involves
302 various predictive approaches, including machine learning, network and response-
303 response models.

304 4 DISCUSSIONS

305 The nuanced terminology usage in the handbook and approaches of adaptation reflect the
306 complex nature of AOP development. Our results suggest several issues that may warrant
307 further discussion, including: 1) a mismatch between the terminology in guideline and the

308 methodology in practice, and 2) considerations of terminology for interdisciplinary
309 communication.

310 **4.1 Mismatch between methodology and terminology**

311 Some putative AOP studies inferred causality of KERs based on predictive models (Gou et
312 al. 2023; Jeong et al. 2018). We argue the use of predictive models in these studies
313 supported the predictive but not the causal aspect of KERs, highlighting a mismatch
314 between the treatment effect framework and predictive models. This mismatch could be
315 demonstrated through possible misinterpretations of probabilistic network models as
316 causal. Probabilistic network models (Koller and Friedman 2009) are based on directed
317 acyclic graphs (DAGs). DAGs are commonly involved in causal reasoning and designing of
318 structural equation modeling (Greenland et al. 1999), e.g. to approximate unbiasedness by
319 quasi-experimental methods. But the use of a DAG is not a guarantee for having a causal
320 model. It depends on context, if a model involving DAGs can provide unbiased average
321 treatment effects for causal inference. In Bayesian modelling, Directed acyclic graphs are
322 used to decompose the joint probability distribution of variables and parameters into a
323 product, with no implication of causality. A Bayesian network is usually referred to a
324 probabilistic network model with binary or categorical nodes. Bayesian networks comprise
325 a wide range of applications, where the network can be inferred from multivariate data,
326 e.g. in machine learning, or given as representative of causal relationships where the
327 probabilities or conditional probability distribution are informed by expert judgment,
328 e.g. in decision analysis. Bayesian networks have been applied to parameterize models
329 where confounding biases are adjusted using quasi-experimental methods such as the

330 backdoor criterion (Pearl 2009) and instrumental variables (Shingaki et al. 2021). With the
331 popularity of Bayesian networks (Kaikkonen et al. 2021; Spinu et al. 2020) comes
332 misconceptions of to what extent a Bayesian network indicates causality, which might
333 come from the use of a DAG. There are studies where results of a Bayesian network were
334 interpreted as causal without employing bias reduction (Chen and Pollino 2012; Jeong et al.
335 2018; Zou et al. 2010). For evaluating the uncertainty in a Bayesian network, beyond the
336 inherent uncertainty embedded within the probability model, it is crucial to consider the
337 process and context in which the Bayesian network is applied (Sahlin et al. 2021).

338 This mismatch is particularly crucial in the stage of putative AOP development, where the
339 assembly of KEs and KERs could be achieved through either causal inference or statistical
340 modeling. The reason is, although predictive inference of KEs and KERs is optional in the
341 recommended best practice for the putative stage (Villeneuve et al. 2014b), an AOP would
342 be incomplete without considerations on the essentiality of KEs, biological plausibility and
343 empirical support of KERs. Therefore, misrepresenting results from predictive models as
344 causal may falsely imply that the dual requirements have been fulfilled in the putative
345 stage, which could risk overlooking causal and observational assessment in the qualitative
346 stage.

347 Furthermore, this mismatch also highlights the importance of integrating predictive
348 models with the evidence from causal or observational inference in the qAOP stage.
349 Current qAOP development often assumes that causality of KERs is well-supported in
350 established AOPs (Burgoon et al. 2020; Gou et al. 2023; Jensen et al. 2023). Yet there are
351 limited approaches to propagate the assumed causality in predictive models. One potential

352 approach is the use of Bayesian methods to integrate known causality of KERs and prior
353 knowledge on KEs in qAOP (Spînu et al. 2022). Another approach involves incorporating
354 causal KERs on the target/ utility/ loss function of predictive models through
355 regularization techniques such as Lasso (Tibshirani 1996). While regularization has been
356 applied in in silico model to mitigate overfitting (Hemmerich and Ecker 2020), there has
357 been limited exploration of its use in adjusting the target/ utility/ loss function on known
358 causality. This may be due to concerns that regularizing on causality may penalize
359 predictive performance. However, it is crucial to re-emphasize that AOPs are only complete
360 with both causal and predictive characteristics (Becker et al. 2015) and good predictive
361 performance alone is insufficient for robust qAOP development. According to the
362 handbook's guideline on the WoE assessment of qAOPs (annex 2), known modulating
363 factors and feedback loops need to be considered along with predictive performance for a
364 qAOP to achieve high level of confidence (Villeneuve et al. 2024).

365 **4.2 Careful considerations of terminology in AOP guidance**

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

369 The use of term "causal" needs to be considered because the frameworks for evaluating
370 causality in the handbook (Villeneuve et al. 2024) and for the use in integrated testing
371 strategies (OECD 2017) may not necessarily conclude causality in the same sense employed
372 in toxicology and generic scientific reasoning. Even the recommended best practice (Becker
373 et al. 2015; Villeneuve et al. 2014b) may be insufficient to achieve unbiased estimates of KE

374 essentiality and KER causality. This is not a critique of current established AOP guidelines
375 and practice because, as shown by our results, the term "causal" is pronounced and utilized
376 in a consistent manner by the handbook, the OECD guidance and studies within the niche of
377 AOP development. Distinctive usage across fields of the same terminology is common.
378 However, adding clarifying annotations on the term "causal" could greatly facilitate
379 interdisciplinary communication and collaboration that potential bias may exist and that
380 the results could be "causal estimates" rather than "causal".

381 To demonstrate, the handbook's framework highlights the inclusion of direct and indirect
382 evidence and a prioritization of direct evidence in the assessment of essentiality of KEs and
383 biological plausibility of KERs. Direct evidence of KE essentiality (handbook section 4B,
384 Table 5) includes evidence from targeted blocking experiments, such as know-out
385 experiments (Skarnes et al. 2011). Knockout experiments pragmatically approximates
386 counterfactual inference and are more robust and reliable than observational evaluation of
387 multiple KEs (see supplementary information). However, results from targeted blocking
388 experiments are not always causal because the caveat of off-target issues applies in design
389 (Höijer et al. 2022). The instrumentation of blocking could be attributable to confounding
390 and uncontrolled mediation on downstream KEs (Skarnes et al. 2011). Unbiased causality
391 could be achieved only if the blocking methods induce blocking in the target KE_{up} but not
392 anything else that causes changes in KE_{down} or their modulating factors. If the instrument
393 to block KE_{up} activates a third, unobserved event that, e.g., increases the activity of KE_{down},
394 it is possible to observe no change in KE_{down}, because the reduction by blocking KE_{up}
395 could be canceled out by the activation from the third event.

396 In targeted knockout experiments, there could be at least two technical reasons leading to
397 the off-target issue. Firstly, the genetic target in knockout procedure is identified by
398 genome-wide association studies, which is association by definition. Confounders may exist
399 since the gene-protein associations are not always one-to-one matched (Olariu et al. 2021).
400 It is well known that the same protein can be coded by different genes or multiple copies of
401 the same gene at different loci (Campenhout et al. 2019). It is also not uncommon to
402 observe the synthesis of multiple proteins from the transcription of one gene. The
403 variational splicing of introns of one gene can be transcripted to different mRNAs, which
404 result in the production of multiple, distinctive proteins (De Conti et al. 2013). Targeted
405 mutation techniques have the potential to become quasi-experimental by implementing
406 bias adjustment methods such as instrumental variables (Landfors et al. 2023). Secondly,
407 assuming the target identification is free from confounding bias, unintended genetic
408 changes could be induced during gene editing (Höijer et al. 2022). Therefore, direct
409 evidence from targeted blocking experiments, as prescribed in the best practice, could be
410 insufficient to guarantee causality.

411 Indirect, non-experimental evidence is evaluated using a WoE framework based on the
412 evolved MOA Hill considerations (Meek et al. 2014). The evolved MOA Hill considerations
413 classifies weights of confidence into high, moderate or low rather than a dichotomous
414 judgement of causality. This is evident that the terms "causal", "causality" or etc. were
415 included only once (not including quotes) in the source paper (Meek et al. 2014) to
416 describe the consideration for KE essentiality that *"if following cessation of repeated*
417 *exposure for various periods, effects are reverse, this constitutes relatively strong evidence*

418 *that key events are causal"*. Furthermore, the adaptations made by the handbook implement
419 no further bias adjustment and focus on the classification of WoE of supporting evidence.

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

422 We also would like to bring up a discussion on the use of the term "essential", particularly
423 in the essentiality of KEs. We argue that adding enhancing annotations on the term
424 "essential" could benefit its practical utility in the assessment of KE essentiality. It is crucial
425 to note that this discussion does not revolve around the distinction between mathematical
426 and toxicological context of "essential". The term "essential" has specialized mathematical
427 context and usage to describe the conditional relationship between two statements. To
428 claim mathematical essentiality of event A for event B, that is to say, if A does not occur,
429 then B cannot occur at all. The validation of mathematical essentiality requires a falsity test.
430 Any case where the occurrence of B without the occurrence of A falsifies essentiality.
431 Obviously, such a falsity test is practically difficult to perform in science (Graesser and Hu
432 2011), especially in biology and toxicology. Most biological events are the consequence of
433 one or several alternative, sufficient causes. The use of "essential" in toxicology is distinct
434 from the mathematical context, including the evolved MOA Hill considerations and AOP
435 development. Under the toxicological context, evaluation of essentiality considers the
436 impact of blocking or reversing KE_{up} on KE_{down} (Meek et al. 2014). A mathematically
437 essential event would cause complete absence in the consequence, while a toxicologically
438 essential event could describe a KE_{up} which, when blocked, prevent or significantly modify
439 KE_{down} (Meek et al. 2014; Villeneuve et al. 2024). The validation of mathematical

440 essentiality of KE_{up} for KE_{down} could be practically unfeasible. The guidelines for AOP
441 development needs to be a pragmatic framework that balances scientific rigor and
442 practicability to serve regulatory and industrial application in the real world. However, the
443 level of significance of modification/ reduction is subjective. Neither the evolved MOA Hill
444 considerations nor the handbook defines the level of modification in KE_{down} to which the
445 changes associated with blocking KE_{up} should be considered as significant and sufficient.
446 Furthermore, the level of modification/ reduction is diverse across systems, organs and
447 molecules by nature. For example, body weight change depends on a myriad of factors,
448 including time period, age, sex and baseline weight. It would be helpful to annotate the
449 handbook's defining question of KE essentiality by adding a description of the level at
450 which the modification in KE_{down} is both statistically significant and biologically sufficient.
451 Further research using the benchmark dose modeling methodology might be useful to
452 characterize the level of modification for KE essentiality.

453 **5 CONCLUSIONS**

454 This study aimed to explore the usage and adaptation of the terms "causal", "observational"
455 and "predictive" for KER description in AOP guidance documents. The results indicate that
456 KERs are described to fulfill both causal and predictive requirements across three stages of
457 AOP development, and distinct approaches are implemented to address different focuses of
458 adaptation across these stages. These findings suggest that the nuanced approach towards
459 the distinction between causal and predictive inference of KERs reveal a mismatch
460 between predictive models and causality in published studies of AOP development,
461 particularly pertaining to the putative stage. Further exploration is needed to integrate

462 causality in qAOP development and application. Further discussions could be helpful to
463 consider the handbook's usage of the terms "causal" and "essential" to facilitate more
464 comprehensive, transparent interdisciplinary communication and practical application of
465 AOP under the academic, regulatory and industrial contexts.

466 **REFERENCE**

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

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

472 Bal-Price A, Crofton KM, Sachana M, Shafer TJ, Behl M, Forsby A, et al. 2015. Putative
473 adverse outcome pathways relevant to neurotoxicity. Critical reviews in toxicology 45:83–
474 91; doi:10.3109/10408444.2014.981331.

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

478 Becker RA, Ankley GT, Edwards SW, Kennedy SW, Linkov I, Meek B, et al. 2015. Increasing
479 scientific confidence in adverse outcome pathways: Application of tailored bradford-hill
480 considerations for evaluating weight of evidence. Regulatory Toxicology and Pharmacology
481 72:514–537; doi:10.1016/j.yrtph.2015.04.004.

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

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

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

491 Brookhart MA, Rassen JA, Schneeweiss S. 2010. Instrumental variable methods in
492 comparative safety and effectiveness research. Pharmacoepidemiology and Drug Safety
493 19:537–554; doi:10.1002/pds.1908.

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

497 Callegaro G, Kunnen SJ, Trairatphisan P, Grosdidier S, Niemeijer M, Hollander W den, et al.
498 2021. The human hepatocyte TXG-MAPr: Gene co-expression network modules to support
499 mechanism-based risk assessment. Arch Toxicol 95:3745–3775; doi:10.1007/s00204-021-
500 03141-w.

501 Campenhout CV, Cabochette P, Veillard A-C, Laczik M, Zelisko-Schmidt A, Sabatel C, et al.
502 2019. Guidelines for optimized gene knockout using CRISPR/Cas9. BioTechniques 66:295–
503 302; doi:10.2144/btn-2018-0187.

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

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

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

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

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

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

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

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

523 Fréchette GR, Sarnoff K, Yariv L. 2022. Experimental economics: Past and future. Annual
524 Review of Economics 14:777–794; doi:https://doi.org/10.1146/annurev-economics-
525 081621-124424.

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

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

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

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

537 Guzelian PS, Victoroff MS, Halmes NC, James RC, Guzelian CP. 2005. Evidence-based
538 toxicology: A comprehensive framework for causation. Human & Experimental Toxicology
539 24:161–201; doi:10.1191/0960327105ht517oa.

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

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

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

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

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

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

554 Höijer I, Emmanouilidou A, Östlund R, Schendel R van, Bozorgpana S, Tijsterman M, et al.
555 2022. CRISPR-Cas9 induces large structural variants at on-target and off-target sites in vivo
556 that segregate across generations. Nature Communications 13:627; doi:10.1038/s41467-
557 022-28244-5.

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

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

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

568 Jeong J, Song T, Chatterjee N, Choi I, Cha YK, Choi J. 2018. Developing adverse outcome
569 pathways on silver nanoparticle-induced reproductive toxicity via oxidative stress in the
570 nematode caenorhabditis elegans using a bayesian network model. Nanotoxicology
571 12:1182–1197; doi:10.1080/17435390.2018.1529835.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

612 PrecisionTox. 2024. Publications-PrecisionTox.

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

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

616 Sager JE, Yu J, Ragueneau-Majlessi I, Isoherranen N. 2015. Physiologically based
617 pharmacokinetic (PBPK) modeling and simulation approaches: A systematic review of
618 published models, applications, and model verification. Drug Metabolism and Disposition
619 43:1823; doi:10.1124/dmd.115.065920.

620 Sahlin U, Helle I, Perepolkin D. 2021. “This is what we don’t know”: Treating epistemic
621 uncertainty in bayesian networks for risk assessment. Integrated Environmental
622 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

644 assessment (IATA). Regulatory Toxicology and Pharmacology 70:629–640;
645 doi:10.1016/j.yrtph.2014.09.009.

646 Villeneuve DL, Angrish MM, Fortin MC, Katsiadaki I, Leonard M, Margiotta-Casaluci L, et al.
647 2018. Adverse outcome pathway networks II: Network analytics. Environmental
648 Toxicology and Chemistry 37:1734–1748; doi:10.1002/etc.4124.

649 Villeneuve DL, Crump D, Garcia-Reyero N, Hecker M, Hutchinson TH, LaLone CA, et al.
650 2014a. Adverse outcome pathway (AOP) development i: Strategies and principles.
651 Toxicological Sciences 142:312–320; doi:10.1093/toxsci/kfu199.

652 Villeneuve DL, Crump D, Garcia-Reyero N, Hecker M, Hutchinson TH, LaLone CA, et al.
653 2014b. Adverse outcome pathway development II: Best practices. Toxicological Sciences
654 142:321–330; doi:10.1093/toxsci/kfu200.

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

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

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