Biological plausibility in environmental health systematic reviews: a GRADE concept paper

Paul Whaley (1,2), Thomas Piggott (3), Rebecca L. Morgan (3), Sebastian Hoffmann (2), Katya Tsaioun (2), Lukas Schwingshackl (4), Mohammed T. Ansari (5), Kristina A. Thayer (6), Holger Schünemann* (3,7,8)

Affiliations

1. Lancaster Environment Centre, Lancaster University, UK

2. Evidence-based Toxicology Collaboration at Johns Hopkins Bloomberg School of Public Health (EBTC)

3. Department of Health Research Methods, Evidence and Impact, McMaster University, 1280 Main St West; Hamilton, ON L8N 3Z5, Canada

4. Institute for Evidence in Medicine, Medical Center - University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany

5. School of Epidemiology and Public Health, University of Ottawa. Room 101, 600 Peter Morand Crescent, Ottawa, Ontario K1G 5Z3, Canada

6. U.S. Environmental Protection Agency (US EPA), Office of Research and Development, Center for Public Health and Environmental Assessment (CPHEA), Chemical Pollutant Assessment Division (CPAD), 1200 Pennsylvania Avenue, NW (8623R), Washington, DC 20460, USA

7. Michael G DeGroote Cochrane Canada and McMaster GRADE Centres; McMaster University, HSC-2C, 1280 Main St West; Hamilton, ON L8N 3Z5, Canada

8. Dipartimento di Scienze Biomediche, Humanitas University, Via Rita Levi Montalcini, 4, 20090 Pieve Emanuele, Milan, Italy,

Abstract

Background: “Biological plausibility” is a concept frequently referred to in environmental and public health when researchers are evaluating how confident they are in the results and inferences of a study or evidence review. Biological plausibility is not, however, a domain of one of the most widely-used approaches for assessing the certainty of evidence (CoE) which underpins the findings of a systematic review, the Grading of Recommendations Assessment, Development and Evaluation (GRADE) CoE Framework. Whether the omission of biological plausibility is a potential limitation of the GRADE CoE Framework is a topic that is regularly discussed, especially in the context of environmental health systematic reviews.

Objectives: We analyse how the concept of “biological plausibility”, as applied in the context of assessing certainty of the evidence that supports the findings of a systematic review, is accommodated under the processes of systematic review and the existing GRADE domains.

Results and Discussion: We argue that “biological plausibility” is a concept which primarily comes into play when direct evidence about the effects of an exposure on a population of concern (usually humans) is absent, at high risk of bias, is inconsistent, or limited in other ways. In such circumstances, researchers look toward evidence from other study designs in order to draw conclusions. In this respect, we can consider experimental animal and *in vitro* evidence as “surrogates” for the target populations, exposures, comparators and outcomes of actual interest. Through discussion of 10 examples of experimental surrogates, we propose that the concept of biological plausibility consists of two principal aspects: a “generalisability aspect” and a “mechanistic aspect”. The “generalisability aspect” concerns the validity of inferences from experimental models to human scenarios, and asks the same question as does the assessment of external validity or indirectness in systematic reviews. The “mechanistic aspect” concerns certainty in knowledge of biological mechanisms and would inform judgements of indirectness under GRADE, and thus the overall CoE. While both aspects are accommodated under the indirectness domain of the GRADE CoE Framework, further research is needed to determine how to use knowledge of biological mechanisms in the assessment of indirectness of the evidence in systematic reviews.

Keywords: systematic review; biological plausibility; surrogates; environmental health; toxicology; epidemiology; Bradford Hill;

Introduction

In environmental and public health research, toxicology, and human health chemical risk assessment (henceforth referred to as “environmental health research”) it is rare to have direct evidence from studies in humans of the effects that environmental exposures might be having on people’s health. This elevates the importance in environmental health research of evidence from experimental animal (*in vivo*) and *in vitro* studies. However, while evidence from *in vivo* and *in vitro* studies has the advantage that exposure can be controlled, the laboratory set-up is only indirectly representative of the human situation which it models - using animals in place of people, artificial cell culture constructs to measure biological processes, and exposure regimens which are often much higher, shorter and more regimented than would be seen in human cases (Rhomberg, 2015).

There is often, therefore, a need to translate the evidence from laboratory experiments to the human scenarios they are informing. Our ability to do this correctly is critical in successfully identifying, quantifying, and limiting health harms from environmental exposures. As systematic reviews become mainstream in environmental health (Bilotta, Milner and Boyd, 2014; Sheehan and Lam, 2015; Morgan *et al.*, 2016; Whaley *et al.*, 2016; Hoffmann *et al.*, 2017), the need for systematic approaches for translating evidence from the laboratory to the human context becomes increasingly important (Lewis et al. 2017).

One concept which is often applied in assessing causality and translating the findings of laboratory experiments to human contexts (or, indeed, one epidemiological context to another) is that of “biological plausibility”. As a concept, biological plausibility was first formalised in 1965 by Sir Austin Bradford Hill, as one of his considerations for establishing causality (Hill, 1965). Bradford Hill argued that the presence of biological plausibility can increase the likelihood that a relationship between an exposure and a health outcome is a causal one. However, despite the evolution of thinking around the concept and the many definitions of “biological plausibility” that are available (see Table 1 for some examples), exactly what constitutes biological plausibility has never been fully or finally characterised. This is particularly true in the context of conducting environmental health systematic reviews. Methodologists, including those in the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group, are frequently challenged by environmental health practitioners about whether and how the assessment of biological plausibility is accommodated in the systematic review process (European Food Safety Authority, 2018).

Source	Definition of "biological plausibility"
Bradford Hill (1965)	"It will be helpful if the causation we suspect is biologically plausible. But this is a feature I am convinced we cannot demand. What is biologically plausible depends upon the biological knowledge of the day."
Wikipedia (Wikipedia contributors, 2014)	"A relationship between a putative cause and an outcome — that is consistent with existing biological and medical knowledge" and "one component of a method of reasoning that can establish a cause-and-effect relationship between a biological factor and a particular disease or adverse event"
European Food Safety Authority (Hardy *et al.*, 2017)	"Consistency between data and biological theory or mechanism"
Last's Dictionary of Epidemiology (International Epidemiological Association, 2001)	The "causal consideration that an observed, potentially causal association between an exposure and a health outcome may plausibly be attributed to causation on the basis of existing biomedical and epidemiological knowledge."
Organisation for Economic Co-operation and Development (OECD, 2016)	Being "consistent with biological knowledge" and "based on extensive previous documentation and broad acceptance"
US Environmental Protection Agency Cancer Guidelines (US Environmental Protection Agency, 2005)	"An inference of causality [which] tends to be strengthened by consistency with data from experimental studies or other sources demonstrating plausible biological mechanisms. A lack of mechanistic data, however, is not a reason to reject causality."

Table 1. Examples of definitions of "biological plausibility"

GRADE and biological plausibility

The GRADE Framework, originally introduced in 2003, is commonly used in public health and healthcare systematic reviews, and increasingly in environmental health (Morgan *et al.*, 2016; Morgan et al. 2019). GRADE contends that assessment of the certainty of evidence for answers to research questions can be successfully operationalised (i.e. conducted accurately, consistently and transparently by different researchers working in different times and places) via systematic consideration of a predefined set of eight "domains" of strengths and limitations of the overall evidence base (Guyatt *et al.*, 2008). The domains that reduce certainty in a body of evidence summarised in a systematic review are risk of bias, inconsistency, indirectness, imprecision and publication bias. The domains which increase certainty are large effect size, presence of a dose-response relationship, and residual opposing confounding (see Figure 1).

These domains are intended to be exhaustive of the concepts necessary for assessing certainty in the evidence, operationalised via a structured reasoning process designed to produce more consistent and transparent results than is achievable by direct application of the considerations of Bradford Hill. Historically, the contention has been that the role played

by assessment of biological plausibility in environmental health assessments is already accommodated either in the GRADE domains or as part of the systematic review process (Schünemann *et al.*, 2011; Hultcrantz *et al.*, 2017). The GRADE Working Group has therefore intentionally not included biological plausibility as a domain in rating the certainty of the evidence.

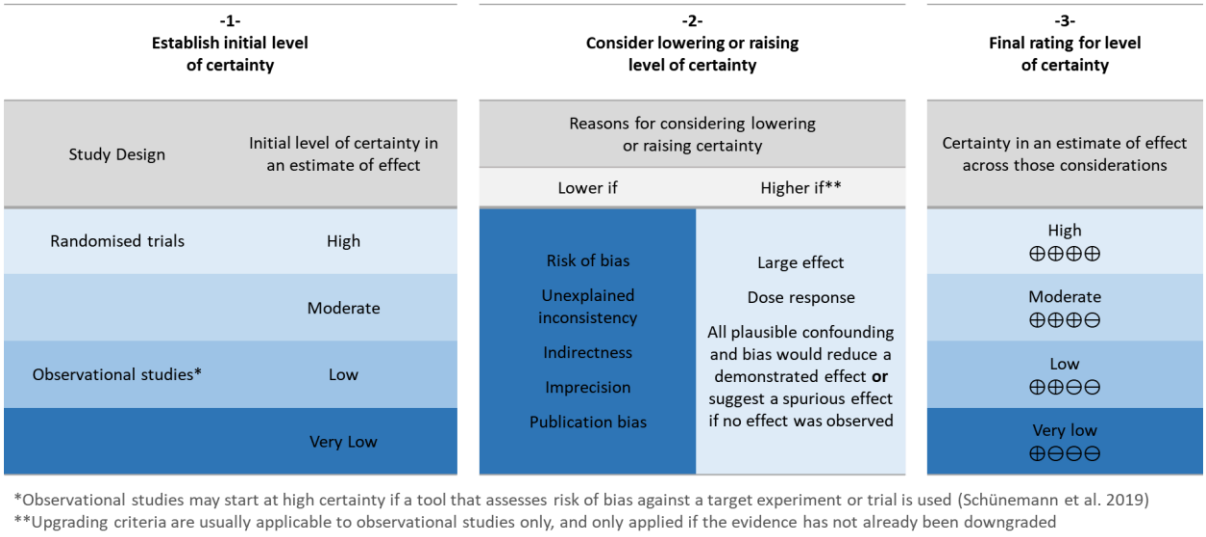

Figure 1. The upgrade and downgrade domains in GRADE and how they are used to determine the overall certainty in evidence for a systematic review. Adapted from Morgan *et al.* (2016).

Our objectives in this paper are as follows: to further elucidate how the systematic review process and the GRADE domains operationalise the assessment of biological plausibility; to describe how the concept of biological plausibility maps onto the process of systematically reviewing environmental health evidence; and answer the question of how "biomedical", "biological", or "epidemiological" knowledge, as referred to in the various definitions of biological plausibility, contributes to rating certainty in a body of evidence summarised in a systematic review.

Our argument consists of five parts. Firstly, we argue that consideration of biological plausibility is not necessary if the body of evidence that is directly reflective of the populations, exposures, comparators and outcomes of concern in a systematic review question is sufficiently certain. Secondly, we note that this situation is rare in environmental health, and that systematic reviews in this field will often need to include indirect evidence from surrogate[1] *in vivo* and *in vitro* experimental models. Thirdly, through 10 examples of the use of surrogates, we show what sort of "biological knowledge" is typically used when researchers are making judgements about biological plausibility.

[1] We define "surrogate" as any property of a study model that is used to estimate the characteristics of a different property. By "property" we mean any controllable or measurable element of study design. This includes population, exposure or intervention, comparator, outcome, and any individual characteristics thereof respectively. For example, rats might be studied in the laboratory as surrogates for human populations, and IQ might be measured as a surrogate for intellectual capacity.

Fourthly, our 10 examples show that the concept of biological plausibility consists of two connected principle aspects, which we call the “generalisability aspect” and the “mechanistic aspect”. The “generalisability aspect” of biological plausibility concerns the extent to which findings from an experimental context apply to a target context of concern. The “mechanistic aspect” concerns certainty in the evidence of biological mechanisms (i.e. the molecular, cellular, and organismal events leading to an outcome). Judgements of the generalisability of a surrogate are informed by evidence of biological mechanisms.

Fifthly, we argue that since the generalisability aspect of biological plausibility and the assessment of indirectness in systematic reviews both concern the external validity of experimental models, it follows that the generalisability aspect of biological plausibility is accommodated under the GRADE domain of indirectness. Insofar as judgements of certainty in biological mechanisms support judgements of the generalisability of a surrogate, then the mechanistic aspect of biological plausibility should also be operationalised under the indirectness domain of GRADE.

We therefore conclude that, while processes and language may be different, the concepts involved in the assessment of biological plausibility are covered by the established domains of GRADE. This means GRADE does not need to introduce additional domains to accommodate biological plausibility. However, we also recognise that GRADE has not yet been applied to the assessment of certainty in a way which takes detailed account of biological mechanisms. We therefore recommend research be conducted to advance understanding of how knowledge about mechanisms should be applied in determining the indirectness of evidence.

We note that we are not providing a complete account of the concept of biological plausibility in all contexts and uses, restricting our focus to its application in the conduct of systematic reviews of exposure-outcome relationships. We also note that there is a potential relationship between biological plausibility and Bradford Hill’s concept of “coherence”. This is acknowledged in argument elsewhere that coherence is covered in GRADE under the domains of inconsistency and indirectness (Schünemann et al., 2011). However, as a different concept to biological plausibility, coherence it is not a focus of this article.

“Biological plausibility” and the inclusion of surrogates in systematic reviews

Systematic review can be defined as the application of methods designed to minimise risk of systematic and random error, and maximise transparency of decision-making, when using existing evidence to answer specific research questions. Asking a specific, focused question is a fundamental step in the systematic review process. Systematic review questions in environmental health are generally characterised in terms of the population, exposure, comparator and outcomes of concern - the PECO mnemonic (Morgan *et al.*, 2018).

One of the principal reasons for characterising environmental health questions and the objectives of systematic reviews in terms of a PECO statement is to facilitate unambiguous

characterisation of the types of studies which will be considered by the authors of a systematic review to be relevant or eligible for answering their question. Studies which are more directly relevant will be of populations, exposures, comparators and outcomes that closely match the PECO of the systematic review; those which are less relevant will match less closely. This concept of fit between a study and the objectives of a systematic review is "external validity" - the extent to which the findings of a study can be generalised to populations, exposures and outcomes outside the context of that study (Higgins JPT et al. (eds), 2019). External validity is part of the indirectness domain in GRADE (Schünemann et al. 2013).

When designing a systematic review, authors need to decide on what their cut-off or threshold for external validity is going to be, i.e. where they draw the line on a study being sufficiently generalisable to their target PECO to be worth including in their review. Where the line is drawn will depend on the review objectives. The way to keep a systematic review relatively small and simple is to define as eligible only those studies whose designs most directly match the PECO characterisation of the systematic review question (see Figure 2A). If the evidence from those studies is sufficiently certain then there is no need to seek out other evidence in support of the findings of the systematic review - a search for indirect evidence need not be undertaken.

A classic example of this scenario, of direct evidence being of high certainty, is smoking causing lung cancer. Several observational studies have investigated doctors (P) who smoke (E), compared them to doctors who do not smoke (C), and assessed the relative risk of lung cancer (O) between the two groups. The studies are at relatively low risk of bias, including confounding; multiple studies of similar design give reasonably consistent results; they are in a representative population; the overall effect size is reasonably precise; there is no evidence that publication bias exaggerates the observed effect size; there is a dose-response relationship; and the effect size is large, with smoking increasing lung cancer risk by a factor of 12-24 (Doll *et al.*, 2005; Pope *et al.*, 2011). These features of the evidence establish with sufficiently high certainty that a causal relationship has been observed, without knowledge of the mechanism by which the exposure causes the outcome.

In such scenarios, the "biological plausibility" of the exposure-outcome relationship does not need to be evaluated - it can be assumed that there must be a discoverable biological mechanism because there is high certainty that the relationship is causal. This is true even when there is little information about the biological mechanism by which the exposure causes its outcome. Conversely, that it is not known why or how the exposure causes the outcome does not undermine certainty that the relationship is causal. This is what we believe Bradford Hill meant when he stated that establishing biological plausibility is helpful but not always necessary for a causal claim (Hill, 1965): "It will be helpful if the causation we suspect is biologically plausible. But this is a feature I am convinced we cannot demand."

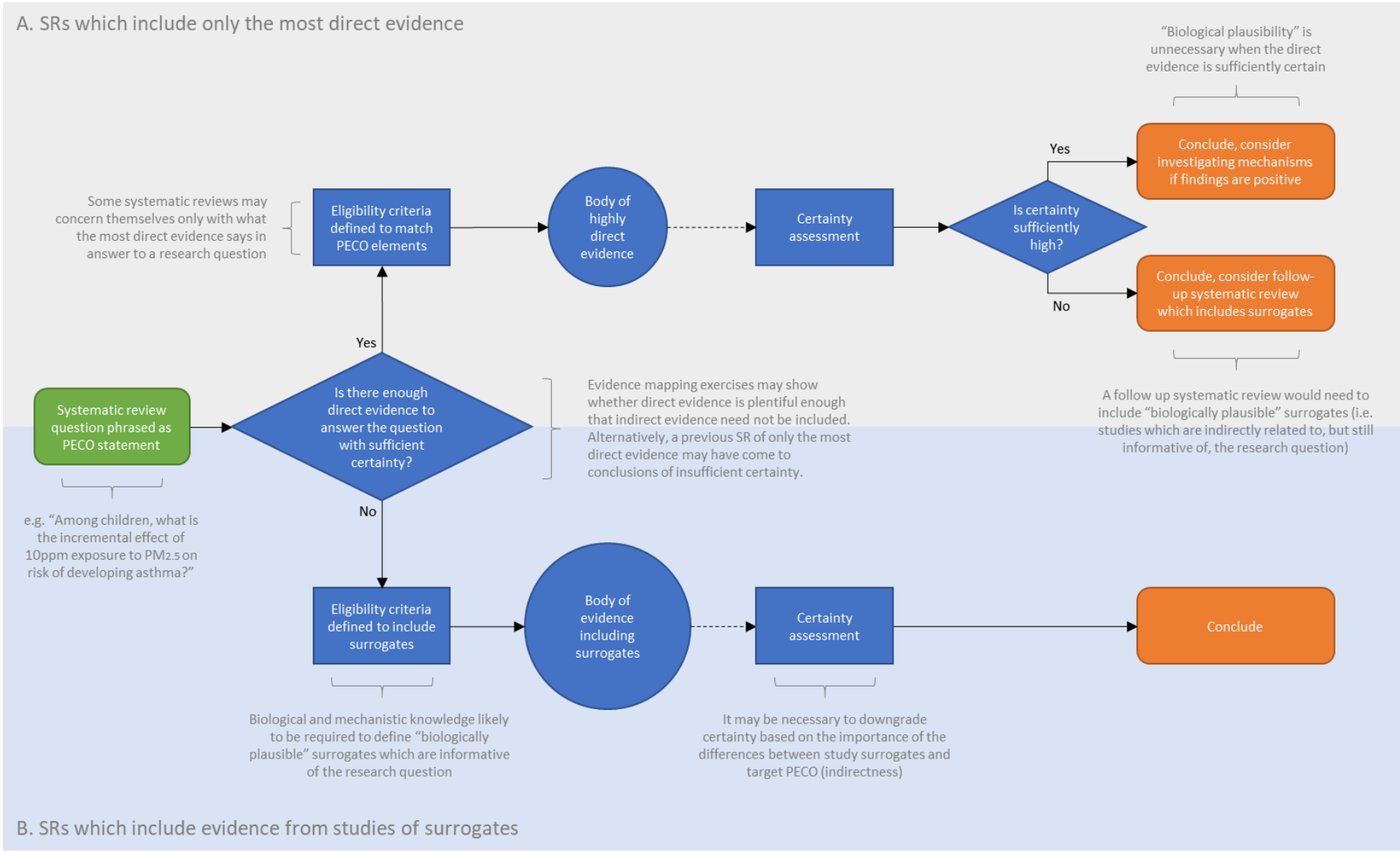

Figure 2. Schematic representation of how it might be decided to include studies of surrogates in a systematic review.

The challenge in environmental health research is that high certainty in direct evidence is a theoretical possibility which is only rarely realised. Usually, environmental health systematic reviews that focus only on the human evidence for a hypothesised exposure-outcome relationship would be expected to yield insufficiently conclusive results, due to the human evidence being highly uncertain or even non-existent. In such circumstances, in order to further investigate and elucidate potential causal relationships between exposures and outcomes, it may become necessary to consider indirect evidence in the form of studies of surrogates (see Figure 2B). This is done in the expectation that including in the systematic review indirect evidence from studies of surrogates will support an assessment of the presence of a causal relationship.

The use of surrogates is familiar in environmental health contexts, which has long been reliant on evidence whereby animal models stand in for target human populations, biomarkers of disease are used in place of observations of clinical health outcomes, and potential health effects of under-studied chemicals are inferred from their similarity to better-researched substances. As with any systematic process, decisions on which surrogates to include in a systematic review should be transparent and well-reasoned, based on evidence of the validity of the decision, and as far as possible defined in advance of conduct of the review (Whaley et al. 2020a). Spurious inclusion of surrogate studies is not just a waste of time and resources: if surrogates are not informative of the question but nonetheless included in the overall analysis, then the validity of the results of the systematic review may be compromised; likewise, spurious exclusion of surrogate studies which should have been included also risks false conclusions.

The “biological plausibility” of choice of surrogates

In conventional environmental health assessments, the consideration of evidence from surrogates is considered to be justifiable insofar as it provides “biologically plausible” support for the hypothesised exposure-outcome relationship in the population of concern (European Food Safety Authority, 2018). In the context of systematic reviews, the GRADE Framework assesses the importance of the indirectness of the surrogate relative to the question being asked. Evidence from surrogates which is too indirect would be excluded from a systematic review; evidence from surrogates which is direct enough to be informative would be included but might be rated down for indirectness (Guyatt *et al.*, 2011).

Here we present 10 examples of the use of surrogates in environmental health assessments. We frame the examples in terms of biological plausibility and describe how the indirectness of the surrogates might be interpreted in the GRADE approach. We then use these examples to show how judgements of biological plausibility map onto the concepts of systematic review. The ten examples and related analyses are summarised in Table 2 and Table 3.

Note that this is a conceptual article describing how ratings of indirectness may be described using the GRADE approach. This information should not be used for decision-making. As for any GRADE concept article, the particular approach described here will require further

validation before it may become official GRADE guidance. Thus, the importance of the indirectness of a surrogate when assessed as part of a systematic review, and therefore any judgements to exclude or downgrade evidence based on its indirectness, may turn out to be different to the judgements which have been made in each of the examples we present.

	Surrogates of higher biological plausibility, for which indirectness is less important	Surrogates of lower biological plausibility, for which indirectness is more important
Population	Animal models for human carcinogenicity of 2-nitropropane	Rat models for human bladder carcinogenicity of saccharine
Exposure (dose)	Extrapolating from high doses to low doses of genotoxic substances	Extrapolating from high doses to low doses of endocrine disrupting chemicals
Exposure (route)	Oral administration of bisphenol-A via gavage, or availability of a pharmacokinetic model to translate intravenous dose to oral equivalent	Intravenous administration of bisphenol-A in absence of pharmacokinetic model to translate intravenous dose to oral equivalent
Exposure (substance)	Inferring estrogenic potential of other bisphenols and from studies of bisphenol-A	Inferring neurotoxicity of organophosphate flame retardants from studies of organophosphate pesticides
Outcome	Maternal serum thyroxine (T4) for child neurodevelopmental outcomes	Biomarkers of Alzheimer's Disease progression in place of clinical measures

Table 2: Summary of the 10 examples used in this manuscript to show how discussion of biological plausibility maps onto the concepts of systematic review.

Surrogate populations

Toxicology has a long history of use of animal models for investigating potential harm to human health from exposure to chemical substances. This is due to the ethical prohibition on conducting experiments in humans that are designed to potentially cause harm, combined with the need for evidence to inform evaluation of chemical health risks, e.g. for regulatory approval and compliance.

One example of where surrogate animal and *in vitro* populations are accepted as providing evidence for health outcomes in human populations of concern is in the assessment of the carcinogenicity of 2-nitropropane. While there is no direct evidence of carcinogenicity in humans, animal and *in vitro* evidence is considered to be sufficiently certain to justify classifying 2-nitropropane as a human carcinogen (Papameletiou *et al.*, 2017). Although the authors did not have a complete account of the mechanism by which 2-nitropropane is a genotoxic carcinogen, they judged it sufficiently biologically plausible that observations in surrogate experimental populations would also be seen in humans that they felt able to draw a conclusion of carcinogenicity.

In a contrasting example, the US Food and Drug Administration (FDA) has deemed evidence from rat models as not relevant for the assessment of saccharin as a bladder carcinogen. This is due to the mechanism by which saccharin causes tumour growth in rats not being present in humans (US National Research Council, 2014). The rat model was originally considered to be predictive but, once the mechanism by which saccharin causes cancer in rats was determined not to be present in humans, the US FDA excluded the rat model from assessment. The US FDA judged the hypothesis that the mechanism by which saccharin causes cancer in rats also occurs in humans as not biologically plausible.

Expressing the reasoning around 2-nitropropane in the concepts of GRADE, it would be to say that surrogate evidence from animal studies is sufficiently direct to be included in a systematic review of the human carcinogenicity of 2-nitropropane. A conclusion about carcinogenicity in humans can be made on the basis of this animal evidence, in spite of its indirectness and a lack of a complete account of the mechanism of carcinogenicity. For saccharin it would be to say that, based on the evidence about the underlying mechanism, the indirectness of the surrogate animal model is unacceptable, i.e. the rat model is too indirect, does not generalise to humans, and therefore is not eligible for inclusion in a systematic review of whether saccharin is a bladder carcinogen.

We again emphasise that the purpose of these examples is to survey how judgements of biological plausibility map onto the concepts and processes of GRADE and systematic reviews. We are not validating any of the judgements that have been made by others in the selected examples. The purpose of the examples is to understand what is involved when researchers are making judgements of biological plausibility, not to determine whether those judgements are valid.

Surrogate outcomes

Surrogate outcomes are used in environmental health research because it is often easier or more ethical in experimental and observational studies to measure biomarkers of disease than clinical outcomes of interest. This is the case when health outcomes may have long latency periods in the population of concern (such as for many cancer types), for particular study designs (e.g. the use of *in vivo* models for allergic contact dermatitis that focus on the induction phase only), or when the observed population may not manifest the apical outcome of interest (e.g. when non-animal test methods address downstream key events relating skin sensitisation).

One example of the use of a surrogate outcome is in a systematic review of the developmental and reproductive toxicity of the biocide triclosan by Johnson et al. (2016). In this case, serum thyroxine concentrations in pregnant women were chosen as a surrogate for the neurodevelopmental health of children. The authors' reasoning was that maternal thyroid hormone levels during pregnancy are predictive of the subsequent neurodevelopmental health of the child - an association described in another systematic review as being "biologically plausible" (Thompson et al. 2018). This can be taken as a judgment by the authors that there is a sufficiently "biologically plausible" relationship between maternal serum thyroxine and neurodevelopment that the former can be treated as a surrogate outcome for the latter.

In contrast, a systematic review of biomarkers for Alzheimer's Disease found insufficient evidence to be able to recommend any biomarker for use as a surrogate outcome for disease progression (McGhee *et al.*, 2014). While it might appear to be "biologically plausible" that Alzheimer's Disease results in specific changes to physical brain structure detectable in an MRI scan (Downey et al. 2017), there seems to be a lack of empirical evidence that directly connects the biomarker to the outcome of concern.

Expressing the triclosan example in the concepts of the GRADE Framework, it would be to say that, in spite of their indirectness, studies which investigate the surrogate outcome can be considered eligible for inclusion in a systematic review of neurodevelopmental toxicity and contribute towards its findings (note that Johnson et al. (2016) used a modified version of the GRADE Framework and did not downgrade for indirectness). For Alzheimer's disease, it would be to say that due to uncertainty around how the surrogate biomarker predicts the ultimate outcome of concern, studies of brain structure biomarkers should either be downgraded more than once for indirectness or excluded from a systematic review if the indirectness is judged to be unacceptable.

Surrogate exposures

Selecting and attributing appropriate importance to surrogate exposures is a complex issue in environmental health systematic reviews. We briefly discuss three aspects of surrogate exposure: route of exposure; administered dose; and active substance. These should provide sufficient illustration of principle, although we note that other aspects of exposure such as measurement of metabolites vs. parent compound, timing of exposure, and other issues, will need consideration in environmental health systematic reviews (Cohen Hubal *et al.*, 2020).

Route

Extrapolating from experimental routes of exposure to the actual routes of exposure likely to be encountered by target populations is a major preoccupation of toxicological risk assessment. For example, toxicology studies which administer bisphenol-A (BPA) to animal test subjects via oral gavage are considered to be of direct relevance to assessing outcomes from dietary exposure. In contrast, intravenous (IV) administration of BPA is typically considered not to be relevant to such assessment, due to the avoidance of first-pass metabolism in the liver (European Food Safety Authority, 2015). However, the relevance of studies using IV administration can increase if knowledge of how BPA is metabolised allows equivalent oral doses to be calculated from IV doses, as this provides what can be interpreted as a "biologically plausible" account of how the two doses are related (Taylor, Welshons and Vom Saal, 2008).

Expressing this in the conceptual framework of GRADE, we would say the indirectness of the route of exposure becomes less important when the exposures of concern can be determined from surrogate exposure routes. Physiologically-based pharmacokinetic (PBPK) models to aid in route-to-route extrapolation are encouraged in chemical assessments (Meek *et al.*, 2013; US Environmental Protection Agency, 2002). The availability of such models may lead to indirect evidence from studies using IV exposure routes being included in a systematic review and potentially rated down fewer levels for indirectness than for scenarios in which such models are unavailable.

Dose

In toxicological research, experiments are often conducted using high doses that are not considered environmentally or occupationally relevant. Many bioassays also merely aim at identifying a maximum tolerated dose of a chemical substance in order to provide a benchmark of toxicity. High dose regimens can raise critical concerns about the indirectness of a study, if the toxicokinetic and toxicodynamic factors by which the administered dose causes an outcome are different from those operating at the dose level of concern (Slikker et al. 2004).

This is a key point of debate about the potential health effects of exposure to endocrine disrupting chemicals: if the administered high dose overwhelms the biological pathway that is involved in the endocrine activity of the active substance, triggering nonspecific pathways that are responsible for the observed outcomes, then there may be critical concerns about the indirectness of the surrogate dose for determining whether the chemical of concern is an endocrine disruptor (Lagarde *et al.*, 2015). An example of this is the causing of endocrine effects via direct damage to the liver (Marty et al. 2018). This would lead to concern that disease induction at high doses via endocrine disruption is not a biologically plausible mechanism, due to differences between the mechanism by which the dose of concern causes the outcome of interest as compared to the mechanism by which the surrogate dose causes the outcome.

In contrast, chemicals which cause cancer by a genotoxic mechanism are considered to operate according to the same mechanism of action at high and low doses (Crump, 1996). In this case, extrapolation from across the dose range is taken to be unproblematic.

Expressing the example of endocrine disruption and genotoxicity in the concepts of the GRADE Framework, the indirectness of a surrogate dose becomes more important when there is evidence of different toxicokinetic and toxicodynamic processes operating at different dose levels. The presence of such differences may result in a decision to exclude evidence from studies using surrogate doses in a systematic review because of unacceptable levels of indirectness. If, on the other hand, it is decided to include the studies that use surrogate doses, rating down for indirectness by two or more levels would be more likely in the example of endocrine disruptors than for genotoxic carcinogens.

Substance

There are many chemicals to which people are potentially exposed which have very few associated toxicology studies. One means for anticipating the potential toxicity of under-studied substances is by extrapolation from evidence of the toxicity of suitably similar chemicals. Often this is based on the demonstration of a common mode of action of toxicity, or sufficient likeness of the surrogate chemical in terms of physical properties that a common mode of action can reasonably be inferred.

For example, the UK Committee on Toxicity (COT) recently evaluated evidence of the neurotoxicity of organophosphate flame retardants (OPFRs) (UK Committee on Toxicity,

2019). Part of their assessment concerned whether the neurotoxicity of OPFRs could be extrapolated from studies of the neurotoxicity of organophosphate pesticides (OPPs). COT determined that OPPs are not a good surrogate exposure for OPFRs, because OPFRs do not inhibit acetylcholinesterase to the same degree as OPPs. COT concluded that there is no "biologically plausible" explanation for how OPPs and OPFRs can cause the same effect, and therefore determined that conclusions about the neurotoxicity of OPFRs should not be derived from evidence of the neurotoxicity of OPPs.

In contrast, since the phase-out of consumer uses of the plastic additive bisphenol-A due to concerns about its potential to act as an oestrogen, considerable research has been conducted into whether replacements such as bisphenol-AF and bisphenol-C may have similar estrogenic potential. Enough similarities in biological effects have been observed for some researchers to suggest that, at least as a group, exposure to some bisphenols may be predictive of the effects of exposure to others (Pelch *et al.*, 2019). Similar suggestions have been made for polyfluorinated compounds (Cousins *et al.*, 2020). When it is more "biologically plausible" that different chemical substances share the same mechanisms by which they exert health effects, then it might be acceptable to use one as a surrogate (also referred to in the environmental health field as an "analogue") for the other.

Expressing the example of OPFRs in the concepts of the GRADE Framework, the absence of explanation for a shared mechanism by which OPPs and OPFRs would exert a neurotoxic effect increases the indirectness of OPFRs as a surrogate for OPPs. If the level of indirectness is unacceptable, it would lead to studies of OPFRs being excluded from a systematic review of their neurotoxicity; if very high, evidence from the surrogate exposure might be included but would be rated down for indirectness, potentially two or three times. For bisphenols and polyfluorinated compounds, if indirectness of surrogates is deemed less important, they may be included in a systematic review and rated down only once for indirectness, or possibly not at all.

Discussion

Biological plausibility as a dual-aspect concept

Our examples show that "biological plausibility" is a concept that can be deployed in multiple scenarios in environmental health assessments. In general, judgements of biological plausibility seem to support judgements of causality insofar as studies of causal relationships in surrogates can be generalised to the target populations, exposures and outcomes of actual concern. These uses extend beyond the definitions of biological plausibility as provided by Bradford Hill and Last's Dictionary of Epidemiology, which define biological plausibility exclusively in terms of biological explanations of a causal relationship between exposure and outcome (see Table 1). When translated into the conceptual underpinnings of GRADE, the uses centre on judging the indirectness of surrogates and describing the impact on certainty in the evidence for the effect an exposure has on a health outcome in a population of concern. These judgements not only govern decisions about the eligibility of

surrogates for a systematic review but also the extent to which a body of evidence based on those surrogates should be downgraded for indirectness.

Our examples also demonstrate that the judgements being made when assessing indirectness are complex. Not only do judgements need to be made about the generalisability of a surrogate, there are also judgements that need to be made about certainty in descriptions of biological mechanism. These judgements are intrinsically connected, as absence of mechanistic explanation limits the ability to generalise from a study surrogate to a target context of concern. This is perhaps most clearly illustrated in the examples of considering whether studies of OPPs are relevant to characterising the potential neurotoxicity of OPFRs, and whether studies of rats are relevant to characterising saccharin as a bladder carcinogen in humans.

Based on these observations, we posit that the concept of “biological plausibility” in fact consists of two principle aspects. We call these the “generalisability aspect” and the “mechanistic aspect”.

We define the *generalisability aspect* of biological plausibility as concerning the validity of generalisations from a surrogate population, exposure, comparator or outcome to a target population, exposure, comparator or outcome of concern, respectively. The generalisability aspect is not about the plausibility of causal claims about the effect of exposures on outcomes, but instead about the extent to which an observation in a surrogate population plausibly generalises to a target population, a surrogate exposure generalises to a target exposure, etc. The generalisability aspect supports judgements of causality insofar as observations are made in studies of surrogates, and the surrogates then generalise to the target contexts of concern.

We define the *mechanistic aspect* as concerning certainty in biological mechanism. Our examples show that judgements of whether a surrogate plausibly generalises to a target context are informed by knowledge of relevant biological mechanisms, i.e. how an exposure causes an outcome in a given biological or experimental system. While this knowledge is not often available, when it is, it has a significant impact on judgements about the generalisability of observations in a surrogate: the higher is the certainty in the knowledge of relevant biological mechanisms (e.g. that similar mechanisms are present in humans and surrogate animal species), the higher is the certainty that a generalisation from a given surrogate to a target context is valid or not. The mechanistic aspect informs the generalisability aspect, as knowledge of mechanism helps determine the validity of generalising from surrogates to target contexts of concern.

These two aspects are different but fundamentally linked: judgements of the plausibility of generalisations are informed by judgements of the plausibility of mechanisms. This connection is illustrated in Figure 3.

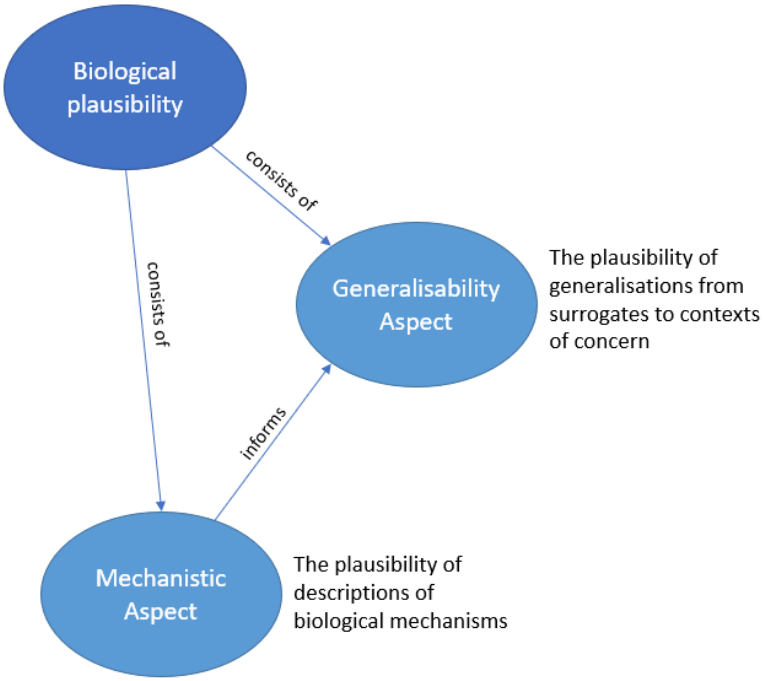

Figure 3. The relationship between the generalisability and mechanistic aspects of biological plausibility.

Biological plausibility and GRADE

How biological plausibility is accommodated within GRADE

Both the generalisability and mechanistic aspects of biological plausibility can be accommodated in the indirectness domain of GRADE. This is because the concept of "generalisability" is the same as the concepts of external validity and indirectness of evidence already familiar in systematic reviews, i.e. the extent to which the results of an experimental or observational study apply to a target context outside of that study (Higgins et al. 2019; Schünemann et al. 2013). The difference is in vocabulary, whereby systematic reviewers talk about the "validity" rather than "plausibility" of a generalisation. Since the generalisability aspect of biological plausibility is asking the same question as the assessment of external validity in systematic reviews, and external validity is subsumed under the GRADE domain of indirectness, it follows that there is no need to extend GRADE to accommodate the generalisability aspect of biological plausibility.

The mechanistic aspect of biological plausibility, because it informs judgements of indirectness or generalisability, is also logically positioned under the indirectness domain of GRADE. This relationship is shown in Figure 4. This means GRADE does not need an additional domain to accommodate assessment of certainty in biological mechanisms. However, this is an interesting category of question for which systematic methods have only recently begun to be explored (Whaley et al. 2020) and we recommend further research on this issue.

Conceptualising matters in this way suggests an operational definition of biological plausibility that maps the concept onto the GRADE framework, as follows: “Biological plausibility is a dual-aspect concept operationalised in systematic reviews as (1) the validity of generalisations from studies of surrogates to target contexts of concern, and (2) certainty in biological mechanisms. Certainty in biological mechanisms informs judgement of the validity of generalisations. When knowledge of biological mechanisms is available, it can have significant impact on judgements of the validity of generalisations.”

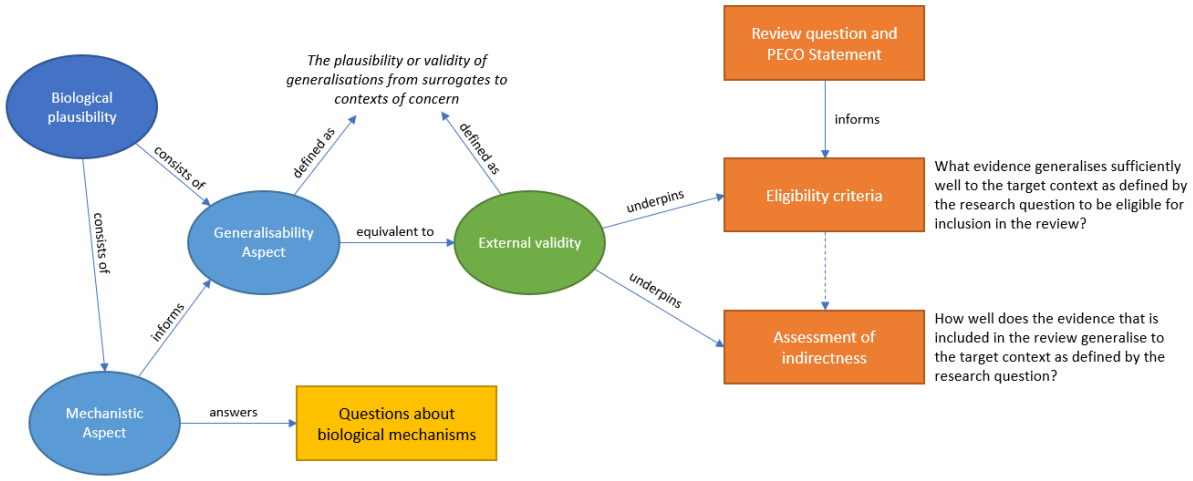

Figure 4. How biological plausibility maps onto the processes of systematic review via the shared concept of external validity, included in GRADE’s indirectness domain. While questions about biological mechanisms (e.g. how an exposure causes an outcome) are independent of a given systematic review, answers to those questions can be highly informative in judging the external validity or indirectness of evidence.

How biological plausibility is accommodated within the processes of systematic review

We have established that biological plausibility maps onto judgements of the generalisability or indirectness of evidence in a systematic review. These judgements are informed by certainty in biological mechanisms. Our task now is to clarify when these judgements are made in the systematic review process. This will show how biological plausibility, in the form of judgements of generalisability informed by knowledge of biological mechanisms, is accounted for when conducting a systematic review. The general principles are articulated below and the specific steps described in Box 1.

Judgements of generalisability or indirectness occur at two stages in systematic reviews, as illustrated in Figure 2. The first stage is in the formulation of eligibility criteria for the inclusion of evidence in a systematic review. Here, the authors decide what study designs are sufficiently generalisable to their question to be worth including in their systematic review. These criteria may be narrowly defined around the most direct evidence, if the authors are attempting to keep their review focused and/or they are confident that looking only at the most direct evidence will provide sufficiently conclusive results. Otherwise, the eligibility criteria may be quite broadly defined, if the authors consider indirect evidence to be of value for their review objectives. Even then, there will be limits to eligibility, as many studies will be so irrelevant to the objective that it would be a waste of time and resources to include them.

Setting these limits, i.e. judging what sort of study designs are informative enough of the research question to be worth including in the systematic review, is a judgement of acceptable indirectness. Mechanistic data, if available, may be of high value in making these judgements.

The second stage is in the judgement of indirectness of the evidence that has been included in the review, when the authors are determining certainty in the evidence on which their results are based. As we show with our example of smoking and lung cancer, in systematic reviews of very direct evidence indirectness is trivial and assessment of biological plausibility unnecessary - the generalisability of findings is a given and mechanistic information is not needed to support judgements of certainty when certainty is already high.

The situation is different for systematic reviews with broadly-defined eligibility criteria. Indirectness in a systematic review with broadly-defined eligibility criteria rapidly becomes very important (due to potentially significant differences between target and surrogate) and complex (due to there being numerous potential differences between the characteristics of the question as formulated in the PECO vs. the included studies). The generalisability of surrogates to the target context of concern is a non-trivial issue and needs to be carefully evaluated. Information about mechanisms is of high value in making these judgements.

How the assessment of biological plausibility is operationalised in systematic reviews of the health effects of environmental exposures which use the GRADE approach for assessing certainty in the evidence

1. Define the systematic review question as a PECO statement: “In population P, what effect does exposure E have on outcome O in comparison to comparator C?”

2. Define as ineligible study models that do not sufficiently generalise to the scenario described in the research question (the generalisability aspect of biological plausibility). These judgements may be informed by knowledge of biological mechanisms (the mechanistic aspect of biological plausibility).

3. Determine the effect of the exposure on the outcome in the studies included in the systematic review. This may require studies to be grouped by design characteristics.

4. Evaluate how well the included evidence generalises to the situation described in the research question for each element of the PECO statement (generalisability aspect). These judgements may be informed by knowledge of biological mechanisms (mechanistic aspect).

5. If it is not certain that the evidence generalises to the research question, rate down the evidence one or more times for indirectness depending on the level of this uncertainty.

Box 1. Explanation of how the concept of biological plausibility is operationalised in the conduct of a systematic review and assessment of certainty in the evidence using the GRADE Framework. Step 2 is the first place where judgements of indirectness may be made. Here, high certainty that an indirect study model does not generalise to the research question may lead to such models being excluded. The US FDA exclusion of rat models from assessments of the bladder carcinogenicity of saccharin is an example of this. Otherwise, systematic reviews of health effects of environmental exposures will likely include indirect study models. (The exception is for deliberately narrowly-focused reviews, as illustrated in Figure 2.) Indirectness judgements are next made in Steps 4 and 5, where the included evidence is assessed under the GRADE domain of indirectness. The UK COT analysis of neurotoxicity of OPFRs based on neurotoxicity of OPPs is an example of when a judgement of lack of certainty in shared biological mechanism results in evidence effectively being rated down for certainty due to indirectness. We note that regulatory frameworks tend to assume a high level of generalisability of a surrogate model unless there is a high level of evidence to the contrary.

Research requirements: judging indirectness at the level of individual studies

The 10 examples in this manuscript show that judgements of external validity are complex and potentially need to be made across multiple related domains of population, exposure, comparator, outcome, and subdomains thereof. Instruments which would facilitate transparent, consistent, and accurate judgements across these domains are not yet available for study-level judgements of indirectness in environmental health and should be developed.

Answering questions about biological mechanisms draws on a wide variety of information about the absorption, distribution, metabolism and excretion ("ADME") of chemical substances, knowledge of mechanisms by which chemicals cause outcomes in both target and observed populations, information about interaction between chemicals and target sites, and the extent to which biomarkers of disease are predictive of clinical outcomes, among many other issues. If mechanistic knowledge is informative of judgements of external validity, and therefore of the indirectness domain in GRADE, it follows that we need to develop methods for assessing certainty in mechanistic knowledge.

Assessing certainty in biological mechanisms would be an important and interesting extension of the GRADE indirectness domain. Unlike questions about associations which are of the form "is X associated with Y?", questions about mechanisms are of the form "how does X cause Y?". Answering this form of question involves describing sequences of biological events, one of which is associated with the next. In principle, event-event associations should be approachable in the same way as exposure-outcome associations, and therefore be amenable to the GRADE approach. A particular challenge we can foresee is in handling the sheer volume of data involved in systematically assessing multiple associated biological events, if very indirect evidence is permitted to enter into the assessment (Whaley et al. 2020b).

We note that developments in the Adverse Outcome Pathway framework may be informative for operationalising the assessment of certainty in biological mechanisms and interpreting indirectness of evidence in systematic reviews (de Vries et al. 2021). Alternatively, the Key Characteristics framework may also provide a structured approach to assessing indirectness via similarity of biological mechanisms (Smith et al. 2016; Guyton et al. 2018).

The 10 examples discussed above give us indications of some the considerations which may reduce concerns about the indirectness of a study included in a systematic review. These are outlined in Table 3 and illustrated, where feasible, in Figure 5. While these are only suggestive selections from the examples we have used in this manuscript, they do illustrate how much of this discussion is already familiar in toxicology and environmental health. This experience should provide a robust platform for further research and should draw on the experience of GRADE and the environmental health communities.

	Potential influencing factors in judging the biological plausibility or external validity of study surrogates
Population	The extent to which the biological pathway connecting exposure to outcome is operating in both the surrogate population and the target population (Figure 5A)
Exposure – dose	The similarity of the toxicodynamic and toxicokinetic processes by which the surrogate dose acts in comparison to that of the dose range of interest
Exposure – route	The similarity by which an organism absorbs and metabolises the substance of concern via the surrogate route as opposed to the target route; or the reliability with which exposure from the surrogate route can be transformed to values which match exposure from the route of interest
Exposure – substance	The extent to which the surrogate molecule influences the biological processes by which the target molecule is thought to elicit its biological effects (Figure 5C)
Outcome	The extent to which a surrogate outcome is predictive of the target outcome of concern (Figure 5B)

Table 3: Summary of potential influencing factors in judging biological plausibility or external validity of study surrogates, as suggested by the examples in this manuscript

Finally, we note that sufficient biological knowledge to permit high-certainty judgements of mechanism and external validity, and thus avoiding rating down for indirectness (and, conversely, being certain that evidence from a surrogate is not relevant), is rare. Absent explanations of mechanism, evidence would either (a) end up being excluded from a systematic review because there is no theoretical route (apart from presumption of relevance) to considering it as eligible, or (b) evidence would be included but its external validity would be unclear, indirectness higher as a consequence, and certainty lower overall. In these cases of low certainty due to unclear external validity of the included studies, significant mechanistic research may be required before it is possible to determine whether one experimental model is more externally valid than another. Currently, in regulatory circumstances where making decisions in the face of uncertainty is important, and mechanistic evidence to support judgements of external validity in a health assessment is limited, the external validity of a choice of surrogate is often assumed unless there is compelling evidence to the contrary (US Environmental Protection Agency, 2005).

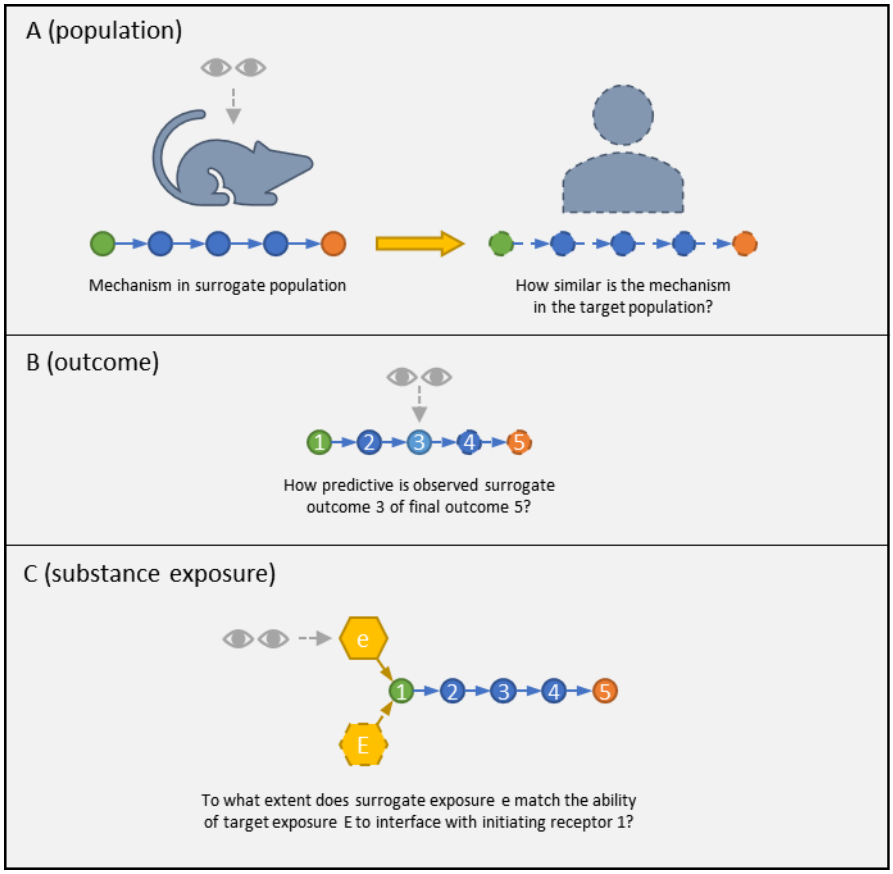

Figure 5: Illustrations of the potential influencing factors in judging biological plausibility or external validity of study surrogates, as suggested by the examples in this manuscript. Circles represent individual events in a biological pathway by which activation of a receptor (green) results in a health outcome (orange). Eyes indicate what surrogate is being observed in an experimental model in lieu of the target of concern (dashed outlines).

Limitations

The attentive reader of our source material will notice that the concept of biological plausibility is rarely clearly applied, even when it appears that it is being discussed. This phenomenon has been observed by other researchers (Dailey, Rosman and Silbergeld, 2018). We have therefore had to impute the concept of biological plausibility to some of our examples - particularly for Alzheimer's Disease, bisphenols, and neurodevelopment - based on surrounding literature and our general understanding of how discussion of biological plausibility is conducted. We believe our imputation to be consistent with use of the concept and intent of the source material, and it is anyway not necessary for the specific term "biological plausibility" to have been used for the concept to have been applied. While a greater number of direct examples could be gathered from a systematic survey of the use of the concept of biological plausibility in the literature, we do not expect that they would invalidate our argument.

Conclusion

We asked what sort of "biomedical", "biological", or "epidemiological" knowledge may influence certainty in the evidence of a systematic review of an exposure-outcome

relationship. Our answer is knowledge of biological mechanisms, which informs judgements of the indirectness of a body of evidence constituted from studies of surrogates. We also set out to determine whether biological plausibility, when applied as a concept relating to certainty in the evidence for the findings of a systematic review, is accommodated by the GRADE domain of indirectness. We have argued that it is, although its full operationalisation will require additional study.

In answering these questions, we have elucidated Bradford Hill's proposition that establishing biological plausibility is helpful but not always necessary for a causal claim (Hill, 1965): "It will be helpful if the causation we suspect is biologically plausible. But this is a feature I am convinced we cannot demand." We have shown that biological plausibility is indeed not necessary for determining that an exposure causes an outcome, so long as the direct evidence for the exposure-outcome relationship is sufficiently certain. The presence of "biological plausibility" can nonetheless be "helpful" to establishing causation. This happens when sufficient information about mechanisms is available to characterise the generalisability of a surrogate, thereby supporting judgements about indirectness of the evidence and potentially permitting a more certain answer to the review question than would be yielded by inclusion of the most direct evidence alone.

Our analysis also broadens the scope of discussion in GRADE of study surrogates. Currently, GRADE guidance only explicitly addresses surrogate outcomes (Guyatt *et al.*, 2011): "Guideline developers should consider surrogate outcomes only when high-quality evidence regarding important outcomes is lacking. When such evidence is lacking […] they should specify the important outcomes and the associated surrogates they must use as substitutes. […] the necessity to substitute the surrogate may ultimately lead to rating down the quality of the evidence because of indirectness." Here, we have extended discussion of eligibility and potential grading of surrogate outcomes to also cover surrogate populations and surrogate exposures.

We have argued that judgements of biological plausibility, at least in their application to determining the relevance of evidence to answering a focused research question, are accommodated under the operational procedures of systematic review and the GRADE domain of indirectness. While vocabulary and processes may differ, we feel confident that there is nothing in biological plausibility that, for this context, is "missing" from GRADE. What is needed, however, are means to operationalise the assessment of the indirectness of included studies and certainty in evidence for biological mechanisms, the outputs of which can be used in determining the extent to which evidence should be rated down for indirectness. Such methods would help bring shape to the amorphous nature of mechanistic evidence and aid in its exploitation in environmental health systematic reviews.

As a final point, we observe a clear parallel between the clinical and public health contexts in which GRADE was developed and the environmental health context in which it is here being applied. The difference is that in clinical contexts, GRADE is nearly always used to evaluate human evidence where treatments are being trialled in people, far downstream from the pre-clinical *in vitro* and animal research that is used to justify conducting a human trial. While

treatments are advanced to human trials based on evidence from preclinical studies, this evidence is often many years old by the time a systematic review is conducted - and therefore preclinical evidence is not needed. In contrast, *in vitro* and *in vivo* research constitutes in many environmental health contexts most of the evidence being dealt with. The fundamental principles for systematically reviewing this evidence are no different to systematic reviews of human evidence, it is just the availability of human evidence that is more limited and mechanisms are often not known. In the context of environmental health, GRADE is, therefore, being applied to a more indirect evidence base which is often focused on events that are further upstream than those dealt with by most healthcare systematic reviews.

Acknowledgements

The authors would like to thank the GRADE Environmental Health Project Group and GRADE Working Group for their contributions to this manuscript, and the Evidence-based Toxicology Collaboration (EBTC) at Johns Hopkins Bloomberg School of Public Health for providing funding to cover the time of PW, KT and SH in working on this manuscript. We would like to thank Dr Andrew Kraft and Dr Michelle Angrish for their technical review of the manuscript, and Dr Daniele Wikoff (ToxStrategies) for detailed discussion of the ideas and concepts we present here. The authors would also thank the European Food Safety Authority and EBTC for organising the Scientific Colloquium, and the participants who contributed to discussions therein, which gave genesis to the concept of this manuscript (European Food Safety Authority, 2018).

Bibliography

Bilotta, G. S., Milner, A. M. and Boyd, I. (2014) 'On the use of systematic reviews to inform environmental policies', *Environmental science & policy*, 42, pp. 67–77. doi: 10.1016/j.envsci.2014.05.010.

Braun, J. M. and Gray, K. (2017) 'Challenges to studying the health effects of early life environmental chemical exposures on children's health', *PLoS biology*, 15(12), p. e2002800. doi: 10.1371/journal.pbio.2002800.

Burns, J. *et al.* (2020) 'Interventions to reduce ambient air pollution and their effects on health: An abridged Cochrane systematic review', *Environment international*, 135, p. 105400. doi: 10.1016/j.envint.2019.105400.

Cohen Hubal, E. A. *et al.* (2020) 'Advancing systematic-review methodology in exposure science for environmental health decision making', *Journal of exposure science & environmental epidemiology*. doi: 10.1038/s41370-020-0236-0.

Cousins, I. T. *et al.* (2020) 'Strategies for grouping per- and polyfluoroalkyl substances (PFAS) to protect human and environmental health', *Environmental science. Processes & impacts*. doi: 10.1039/d0em00147c.

Crump, K. S. (1996) 'The linearized multistage model and the future of quantitative risk

assessment', *Human & experimental toxicology*, 15(10), pp. 787–798. doi: 10.1177/096032719601501001.

Dailey, J., Rosman, L. and Silbergeld, E. K. (2018) "Evaluating biological plausibility in supporting evidence for action through systematic reviews in public health," *Public health*. Elsevier BV, 165, pp. 48–57. doi: 10.1016/j.puhe.2018.08.015.

de Vries et al. (2021) "Applying evidence-based methods to the development and use of adverse outcome pathways", *ALTEX - Alternatives to animal experimentation*, 38(2), pp. 336-347. doi: 10.14573/altex.2101211

Doll, R. *et al.* (2005) 'Mortality from cancer in relation to smoking: 50 years observations on British doctors', *British journal of cancer*, 92(3), pp. 426–429. doi: 10.1038/sj.bjc.6602359.

Downey, A. *et al.* (2017) 'communicating with the public about interventions to prevent cognitive decline and dementia', in *Preventing Cognitive Decline and Dementia: A Way Forward*. National Academies Press (US).

European Food Safety Authority (2015) 'Scientific Opinion on the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs: PART II - Toxicological assessment and risk characterisation', 13(1). doi: 10.2903/j.efsa.2015.3978.

European Food Safety Authority (2018) 'EFSA Scientific Colloquium 23 – Joint European Food Safety Authority and Evidence-Based Toxicology Collaboration Colloquium Evidence integration in risk assessment: the science of combining apples and oranges 25–26 October 2017 Lisbon, Portugal', *EFSA Supporting Publications*, 15(3). doi: 10.2903/sp.efsa.2018.EN-1396.

Gauderat, G. *et al.* (2017) 'Prediction of human prenatal exposure to bisphenol A and bisphenol A glucuronide from an ovine semi-physiological toxicokinetic model', *Scientific reports*, 7(1), p. 15330. doi: 10.1038/s41598-017-15646-5.

Guyatt, G. H. *et al.* (2008) 'GRADE: an emerging consensus on rating quality of evidence and strength of recommendations', *BMJ* , 336(7650), pp. 924–926. doi: 10.1136/bmj.39489.470347.AD.

Guyatt, G. H. *et al.* (2011) 'GRADE guidelines: 8. Rating the quality of evidence--indirectness', *Journal of clinical epidemiology*, 64(12), pp. 1303–1310.

Guyton et al. (2018) Application of the key characteristics of carcinogens in cancer hazard identification, *Carcinogenesis*, Volume 39, Issue 4, April 2018, Pages 614–622, https://doi.org/10.1093/carcin/bgy031

Hardy, A. *et al.* (2017) 'Guidance on the use of the weight of evidence approach in scientific assessments', *EFSA Journal*, 15(8). doi: 10.2903/j.efsa.2017.4971.

Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (ed.) (2019) *Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019)*. Cochrane. Available at: www.training.cochrane.org/handbook.

Hill, A. B. (1965) 'The Environment and Disease: Association or Causation?', *Proceedings of the Royal Society of Medicine*, 58, pp. 295–300. Available at: https://www.ncbi.nlm.nih.gov/pubmed/14283879.

Hoffmann, S. *et al.* (2017) 'A primer on systematic reviews in toxicology', *Archives of toxicology*, 91(7), pp. 2551–2575. doi: 10.1007/s00204-017-1980-3.

Hultcrantz, M. *et al.* (2017) 'The GRADE Working Group clarifies the construct of certainty of evidence', *Journal of clinical epidemiology*, 87, pp. 4–13. doi: 10.1016/j.jclinepi.2017.05.006.

International Epidemiological Association (2001) *A Dictionary of Epidemiology*. Oxford University Press. Available at: https://play.google.com/store/books/details?id=nQmhQgAACAAJ.

Johnson, P. I. *et al.* (2016) 'Application of the Navigation Guide systematic review methodology to the evidence for developmental and reproductive toxicity of triclosan', *Environment international*. The Authors, 92-93, pp. 716–728. doi: 10.1016/j.envint.2016.03.009.

Lagarde, F. *et al.* (2015) 'Non-monotonic dose-response relationships and endocrine disruptors: a qualitative method of assessment', *Environmental health: a global access science source*, 14, p. 13. doi: 10.1186/1476-069X-14-13.

Lewis, S. J. *et al.* (2017) 'Developing the WCRF International/University of Bristol Methodology for Identifying and Carrying Out Systematic Reviews of Mechanisms of Exposure–Cancer Associations', *Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 26(11), pp. 1667–1675.

Lin, J. H. (2008) 'CSF as a surrogate for assessing CNS exposure: an industrial perspective', *Current drug metabolism*, 9(1), pp. 46–59. doi: 10.2174/138920008783331077.

McGhee, D. J. M. *et al.* (2014) 'A systematic review of biomarkers for disease progression in Alzheimer's disease', *PloS one*, 9(2), p. e88854. doi: 10.1371/journal.pone.0088854.

Meek, M. E. B. *et al.* (2013) 'Case study illustrating the WHO IPCS guidance on characterization and application of physiologically based pharmacokinetic models in risk assessment', *Regulatory toxicology and pharmacology: RTP*, 66(1), pp. 116–129. doi: 10.1016/j.yrtph.2013.03.005.

Morgan, R. L. *et al.* (2016) 'GRADE: Assessing the quality of evidence in environmental and occupational health', *Environment international*. Elsevier Ltd, 92-93, pp. 1–6. doi: 10.1016/j.envint.2016.01.004.Morgan, R. L. *et al.* (2018) 'Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes', *Environment international*. Elsevier, (July), pp. 1–5. doi: 10.1016/j.envint.2018.07.015.

Morgan, R. L. *et al.* (2019) 'GRADE guidelines for environmental and occupational health: A new series of articles in Environment International', *Environment international*, 128, pp. 11–12.

OECD (2016) 'Users' Handbook supplement to the Guidance Document for developing and assessing Adverse Outcome Pathways', *Env/Jm/Mono(2016) 12*, (OECD Series on Adverse Outcome Pathways1), p. 63. doi: 10.1787/5jlv1m9d1g32-en.

Papameletiou, D. *et al.* (2017) 'SCOEL/REC/300 2-Nitropropane - Recommendation from the Scientific Committee on Occupational Exposure Limits'. Directorate-General for Employment, Social Affairs and Inclusion (European Commission) , Scientific Committee on Occupational Exposure Limits. doi: 10.2767/841951.

Pelch, K. E. *et al.* (2019) 'Characterization of Estrogenic and Androgenic Activities for Bisphenol A-like Chemicals (BPs): In Vitro Estrogen and Androgen Receptors Transcriptional Activation, Gene Regulation, and Binding Profiles', *Toxicological sciences:*

an official journal of the Society of Toxicology. doi: 10.1093/toxsci/kfz173.

Pope, C. A., 3rd *et al.* (2011) 'Lung cancer and cardiovascular disease mortality associated with ambient air pollution and cigarette smoke: shape of the exposure-response relationships', *Environmental health perspectives*, 119(11), pp. 1616–1621. doi: 10.1289/ehp.1103639.

Prozialeck, W. C. (2013) 'Biomarkers for Cadmium', in Kretsinger, R. H., Uversky, V. N., and Permyakov, E. A. (eds) *Encyclopedia of Metalloproteins*. New York, NY: Springer New York, pp. 272–277. doi: 10.1007/978-1-4614-1533-6_33.

Prozialeck, W. C. and Edwards, J. R. (2010) 'Early biomarkers of cadmium exposure and nephrotoxicity', *Biometals: an international journal on the role of metal ions in biology, biochemistry, and medicine*, 23(5), pp. 793–809. doi: 10.1007/s10534-010-9288-2.

Radke, E. G. *et al.* (accepted) 'Application of US EPA IRIS systematic review methods to the health effects of phthalates: lessons learned and path forward', *Environment international*.

Rhomberg, L. (2015) 'Hypothesis-Based Weight of Evidence: An Approach to Assessing Causation and its Application to Regulatory Toxicology', *Risk analysis: an official publication of the Society for Risk Analysis*, 35(6), pp. 1114–1124. doi: 10.1111/risa.12206.

Schünemann, H. *et al.* (2011) 'The GRADE approach and Bradford Hill's criteria for causation', *Journal of Epidemiology & Community Health*, 65(5), pp. 392–395. doi: 10.1136/jech.2010.119933.

Schünemann, H. J. *et al.* (2013) 'Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions', *Research synthesis methods*, 4(1), pp. 49–62.

Schünemann, H. J. *et al.* (2019) 'GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence', *Journal of clinical epidemiology*, 111, pp. 105–114.

Sheehan, M. C. and Lam, J. (2015) 'Use of Systematic Review and Meta-Analysis in Environmental Health Epidemiology: a Systematic Review and Comparison with Guidelines', *Current environmental health reports*, 2(3), pp. 272–283. doi: 10.1007/s40572-015-0062-z.

Slikker, W., Jr *et al.* (2004) 'Dose-dependent transitions in mechanisms of toxicity: case studies', *Toxicology and applied pharmacology*, 201(3), pp. 226–294.

Smith et al. (2016) 'Key characteristics of carcinogens as a basis for organizing data on mechanisms of carcinogenesis. *Environ Health Perspect* 124:713–721; http://dx.doi.org/10.1289/ehp.1509912

Marty, M. S. *et al.* (2018) 'Distinguishing between endocrine disruption and non-specific effects on endocrine systems', *Regulatory toxicology and pharmacology: RTP*, 99, pp. 142–158.

Taylor, J. A., Welshons, W. V. and Vom Saal, F. S. (2008) 'No effect of route of exposure (oral; subcutaneous injection) on plasma bisphenol A throughout 24h after administration in neonatal female mice', *Reproductive toxicology* , 25(2), pp. 169–176. doi: 10.1016/j.reprotox.2008.01.001.

Thompson, W. *et al.* (2018) 'Maternal thyroid hormone insufficiency during pregnancy and risk of neurodevelopmental disorders in offspring: A systematic review and meta-analysis',

Clinical endocrinology, 88(4), pp. 575–584.

UK Committee on Toxicity (COT) (2019) 'Statement on phosphate-based flame retardants and the potential for neurodevelopmental toxicity'. Available at: https://cot.food.gov.uk/cotstatements/cotstatementsyrs/cot-statements-2019/cot-phosphate-based-flame-retardants-statement.

US Environmental Protection Agency (2002) 'A review of the reference dose and reference concentration processes'. Report Number EPA/630/P-02/002F. Washington, DC. Available at: https://www.epa.gov/sites/production/files/2014-12/documents/rfd-final.pdf

US Environmental Protection Agency (2005) 'Guidelines for Carcinogen Risk Assessment'. Available at: https://www.epa.gov/risk/guidelines-carcinogen-risk-assessment.

US National Research Council (2014) *Review of EPA's Integrated Risk Information System (IRIS) Process*. The National Academies Press. Available at: http://www.nap.edu/openbook.php?record_id=18764.

Whaley, P. *et al.* (2016) 'Implementing systematic review techniques in chemical risk assessment: Challenges, opportunities and recommendations', *Environment international*. The Authors, 92-93, pp. 556–564. doi: 10.1016/j.envint.2015.11.002.

Whaley, P. *et al.* (2020a) 'Recommendations for the conduct of systematic reviews in toxicology and environmental health research (COSTER)', *Environment international*, 143, p. 105926.

Whaley, P. *et al.* (2020b) 'Knowledge Organization Systems for Systematic Chemical Assessments', *Environmental health perspectives*, 128(12), p. 125001.

Wikipedia contributors (2014) *Biological plausibility*, *Wikipedia, The Free Encyclopedia*. Available at: https://en.wikipedia.org/w/index.php?title=Biological_plausibility&oldid=614374435 (Accessed: 16 October 2019).