Practical challenges in assessing indirectness and the implications for integrating multiple streams of evidence in systematic reviews

Paul Whaley
Evidence Based Toxicology Collaboration Research Fellow
Lancaster Environment Centre, UK
National Academy of Sciences, Engineering and Medicine, Washington DC
Evidence Integration Workshop, 3 June 2019

Lancaster Environment Centre | Lancaster University | ebtc | Johns Hopkins Bloomberg School of Public Health

About me

- Researcher at Lancaster University and the Evidence-Based Toxicology Collaboration at Johns Hopkins Bloomberg School of Public Health
- Editor for Systematic Reviews at *Environment International* (IF 7.297)
- Focus on systematic review methods for environmental health research: frameworks for systematic evidence surveillance and synthesis; critical appraisal tools; research standards; quality assurance and control in SR publishing

Today's themes

- Systematic review as a grounded approach to evidence review
- A PECO-based framework for assessing external validity of studies
- Evidence that successfully grounding SRs is extremely challenging
- How our PECO framework anticipates a computational approach to SR
- Research needs for delivering grounded, computational SRs

Recap of systematic review and evidence integration

A PECO-based framework for evidence integration

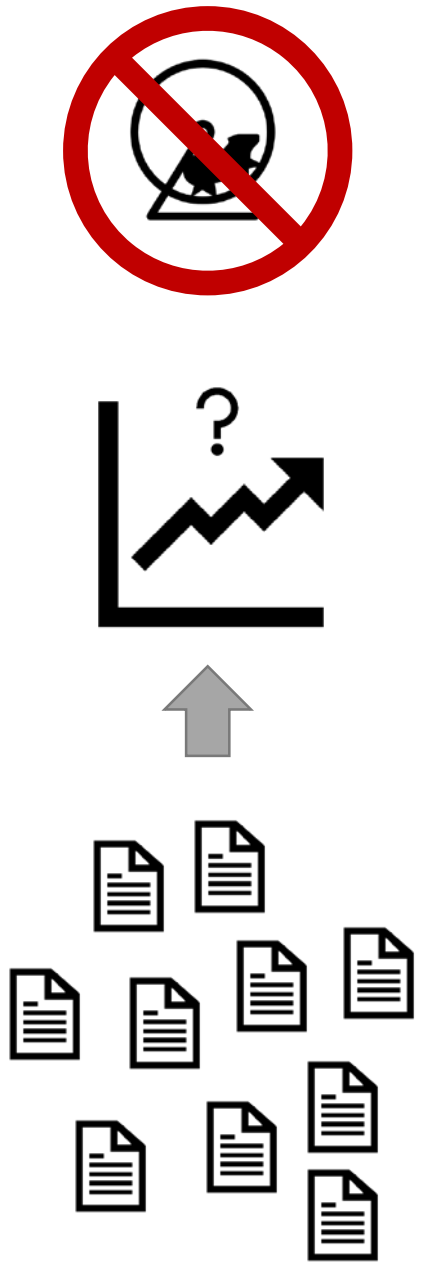
Practical challenges in achieving grounded analysis

Solution: a computational approach

Conclusions and credits

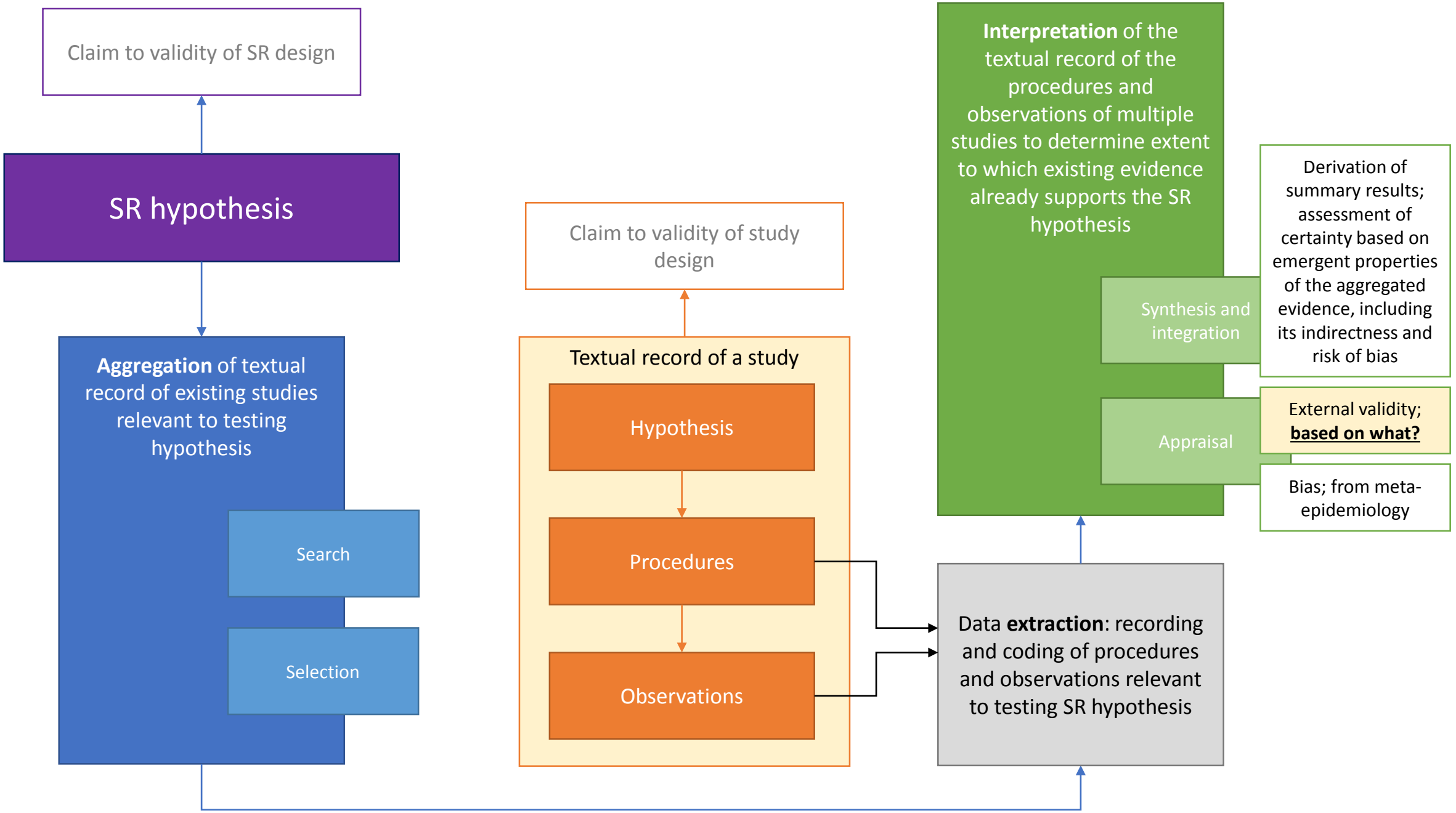
What is a systematic review?

- A systematic review is a research project which tests a hypothesis using pre-existing evidence instead of conducting a novel experiment
- The test should minimise bias introduced by (a) the evidence included in the review, and (b) by the performance of the review
  - Include all the evidence relevant to testing the hypothesis (search and screening)
  - Appraise the quality of the evidence (at level of individual study and body of evidence)
  - Synthesise the evidence into a summary result (qualitative & quantitative methods)



Systematic review = grounded interpretation

- SR is an advance on traditional narrative review because it uses explicit, discussable methods to **ground** the test of the hypothesis
- SRs are grounded when they connect interpretation of the validity of study procedures and observations with:
  a. the textual record in the study documents of those procedures & observations
  b. empirical evidence of the validity of the procedures described in that record
- Can't take grounding for granted, but because SR methods are explicit, they can be repeated, evaluated and deliberately changed


Claim to validity of SR design
SR hypothesis
Aggregation of textual record of existing studies relevant to testing hypothesis
Search
Selection
Claim to validity of study design
Textual record of a study
Hypothesis
Procedures
Observations
Data extraction: recording and coding of procedures and observations relevant to testing SR hypothesis
Interpretation of the textual record of the procedures and observations of multiple studies to determine extent to which existing evidence already supports the SR hypothesis
Synthesis and integration
Derivation of summary results; assessment of certainty based on emergent properties of the aggregated evidence, including its indirectness and risk of bias
Appraisal
External validity; based on what?
Bias; from meta-epidemiology

What is “evidence integration”?

- Evidence integration is based on a concept of dividing evidence into streams (or lines) of readily-comparable populations – usually animal vs. human, though could be a species, genus, or family
- Evidence is synthesised to produce summary results of effect of exposure in each stream
- Certainty of the evidence for the effect is assessed for each stream
- Integration is a function of combined certainty across each stream, generating a judgement of the overall level of evidence
- In the OHAT framework, mechanistic data can inform changes to the level of evidence; in the 2019 update to the IARC preamble, mechanistic evidence is a distinct stream in its own right

Integrating mechanistic information in SRs

- Current approaches were designed to support qualitative hazard classification, not obviously applicable to complex analysis objectives (e.g. quantifying health effects of exposures)
- We already exclude or combine multiple study designs according to principles of relevance or similarity which are informed by mechanistic data
- Mechanistic studies are conducted because they describe and/or predict health outcomes in a target population – why separate them from the whole-organism models of which they are intended to be informative?
- Can we do more to systematically incorporate mechanistic evidence into systematic reviews of exposures?

Recap of systematic review and evidence integration

A PECO-based framework for evidence integration

Practical challenges in achieving grounded analysis

Solution: a computational approach

Conclusions and credits

The role of PECO statements in SRs

- SR = test of a hypothesis using existing evidence
- Hypothesis interpreted as a research question, formulated as a Population-Exposure-Comparator-Outcome statement
- Common research scenario in environmental health: there is a suspected relationship between an exposure and an outcome, but the nature of the relationship is unknown (scenario 1, right)

P: Among adult females, what is the effect of
E: 1 µg/kg bw childhood organochlorine levels in blood, versus
C: 1 µg/kg bw incremental increase on
O: endometriosis?

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Preface

Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes

Rebecca L. Morgan[a], Paul Whaley[b], Kristina A. Thayer[c], Holger J. Schünemann[a,d,*]

[a] Department of Health Research Methods, Evidence, and Impact (Formerly the Department of Clinical Epidemiology & Biostatistics) & Michael G. DeGroote Cochrane Canada Centre, McMaster University, Health Sciences Centre, Room 2C14, 1280 Main Street West, Hamilton, ON L8S 4K1, Canada
[b] Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK
[c] Integrated Risk Information System (IRIS) Division, National Center for Environmental Assessment (NCEA), Office of Research and Development, US Environmental Protection Agency, Building B (Room 2111), Research Triangle Park, NC 27711, USA
[d] Department of Medicine, McMaster University, Health Sciences Centre, Room 2C14, 1280 Main Street West, Hamilton, ON L8S 4K1, Canada

ARTICLE INFO

Handling Editor: Paul Whaley

1. PECO formulation guidance

A clearly-framed question creates the structure and delineates the approach to defining research objectives, conducting systematic reviews and developing health guidance (Guyatt et al., 2011; Armstrong et al., 2007). To assess the association between exposures and outcomes, including in the field of nutrition, environmental and occupational health, the concept of defining the Population (including animal species), Exposure, Comparator, and Outcomes (PECO) as pillars of the question is increasingly accepted (Morgan et al., 2016; Morgan et al., n.d.). Thus, the PECO defines the objectives of the review or guideline. Furthermore, the PECO informs the study design or inclusion and exclusion criteria for a review, as well as facilitating the interpretation of the directness of the findings based on how well the actual research findings represent the original question.

Previously, we have recognized the importance of PECOs for directing the assessments of benefits and harms, identification of ex-

(Guyatt et al., 2011). The Cochrane Collaboration emphasizes the importance of a well-formulated research question to guide an intervention review and provides clarity about the individual PICO components (Higgins and Green, 2011); however, a review of 313 research studies reported that over half (54%) of the studies did not report on the four PICO components (Thabane et al., 2009).

In environmental, public and occupational health research, specific challenges exist with identifying the exposure and comparator within the PECO. In fact, in these fields there are fundamental differences to formulating questions about interventions and comparators in the PICO framework (Guyatt et al., 2011). The Cochrane Handbook, widely recognized as reference guide for systematic reviews, does not specifically address the development of questions for *reviews of exposures* (Higgins and Green, 2011). Other organizations have reported adapting PICO to PECO for studies of unintentional exposure (Collaboration for Environmental Evidence, 2013; NTP (National Toxicology Program), 2015; Woodruff and Sutton, 2014). For example, the Collaboration for

Including indirect evidence

- Necessary in a SR of an exposure-outcome relationship when we do not have certain evidence within the strict confines of the PECO
- Look at intermediate outcomes, disease markers, animal models, similar chemicals (read-across), etc. etc.
- All indirect evidence but still relevant to the question, and therefore could increase certainty in test of hypothesis
- There are lots of ways in which this evidence can be organised

Example: Matta et al. (2019)

Table 2
Body of evidence structure based on major experimental outcomes of endometriosis to guide grouping endpoints and experiments.

Level	Endometriosis-related outcomes	Endpoint/assay examples	Body of evidence grouping examples
Primary/apical outcomes	Spontaneous endometriosis	***In vivo***: onset after chronic/transgenerational exposure in non-human primates	1-Spontaneous endometriosis in animals
	Migration/attachment	***In vivo***: experiments evaluating the invasiveness of implants in rodents or primates ***In vitro***: migration assays in cell models	2- Invasiveness of endometriotic tissue in animals 3 – Invasiveness of endometriotic tissue in cell cultures
	Survival/proliferation/ apoptosis	***In vivo***: experiments on proliferation/expansion of endometriotic lesions in rodents and/or primates ***In vitro***: proliferation/viability/apoptosis cell assays	4 – Survival/proliferation of lesions in animals 5- Proliferation in cell culture
Intermediary /secondary	Progesterone resistance	***In vivo***: PR-B/A expression	6- Progesterone resistance in animals
	Aromatase/steroidogenic pathway	***In vitro***: CYP19A1 expression	7- Disruption of aromatase pathway in cell culture
	Inflammatory cytokines	***In vivo***: IL6 levels	8 - Inflammation in animals
	Other outcomes: immunosuppression, oxidative stress		

Interpreting Matta et al. into PECOs

Level	Population	Exposure	Comparator	Outcome
1°	Non-human primate	Chronic	?	Spontaneous endometriosis
	Non-human primate with implanted tissue	Transgen	?	Invasiveness of implanted tissue
	Rodent	Chronic	?	Proliferation of endometriotic tissue
2°	In vivo	?	?	PR-B/A expression (progesterone resistance)
	In vitro	?	?	CYP19A1 expression (aromatase pathway)
	In vivo	?	?	Inflammation

- As described, relationship between included studies, hypotheses under test and the relevant PECOs are ambiguous – characteristics need to be more tightly defined
- In actuality, we probably don't need to define in advance all the potentially relevant sub-PECOs (cumbersome, p-hacking) – can't we just observe how direct the evidence is?

Proposal: PECOs as a directness framework

Relative to the PECO which is the target of a SR, all evidence is to some extent indirect, and may therefore be evaluated as follows:

- Define the target PECO (tPECO) for the SR, as we do already
- Extract the experimental PECO (ePECO) from each included study
- Evaluate the similarity of each ePECO to the SR tPECO (ePECO→tPECO)
- Describe directness of the evidence overall as a function of how the ePECOs map in aggregate onto the SR tPECO

What this might look like…

	P features				E features			C features	O features
Study	**Specie**	**L. Org.**	**Age**	**Sex**	**Chem**	**Dose**	**Timing**	**Dose**	**Outcome**
Target	Human	Whole organism	Pre-menopause	Female	OC	1 µg/kg bw	Pre-puberty	1 µg/kg bw increments	Endometriosis
Ref013	Human	Whole organism	Adult	Female	Furan mix	High exposure group	Up to 16 years age	Low exposure group	Endometriosis
Ref852	Human	HESC cells	-	Female	TCDD	10uM solution	-	10 uM increments	Migration
Ref134	Wistar Rat	Whole organism	24 months	Male	Chlorpyrifos	1000 µg/kg bw/d	Until weaning	Vehicle	PR-B/A expression

- Allows us to describe all types of study design using the same set of categories
- We can make comparisons between experimental PECOs and our target question, without having to divide evidence up into streams beforehand
- Makes explicit the information being interpreted (if not yet the rules for interpretation)

	P features				E features			C features	O features
Study	**Specie**	**L. Org.**	**Age**	**Sex**	**Chem**	**Dose**	**Timing**	**Dose**	**Outcome**
Target	Human	Whole organism	Pre-menopause	Female	OC	1 µg/kg bw	Pre-puberty	1 µg/kg bw increments	Endom
Ref013	Human	Whole organism	Adult	Female	Furan mix	High exposure group	Up to 16 years age	Low exposure group	Endometriosis
Ref852	Human	HESC cells	-		TCDD	10uM solution	-	10 uM increments	Migration
Ref134	Wistar Rat	Whole organism	24 months	Male	Chlorpyrifos	1000 µg/kg bw/d	Until eaning	Vehicle	PR-B/A expression

A

B

C

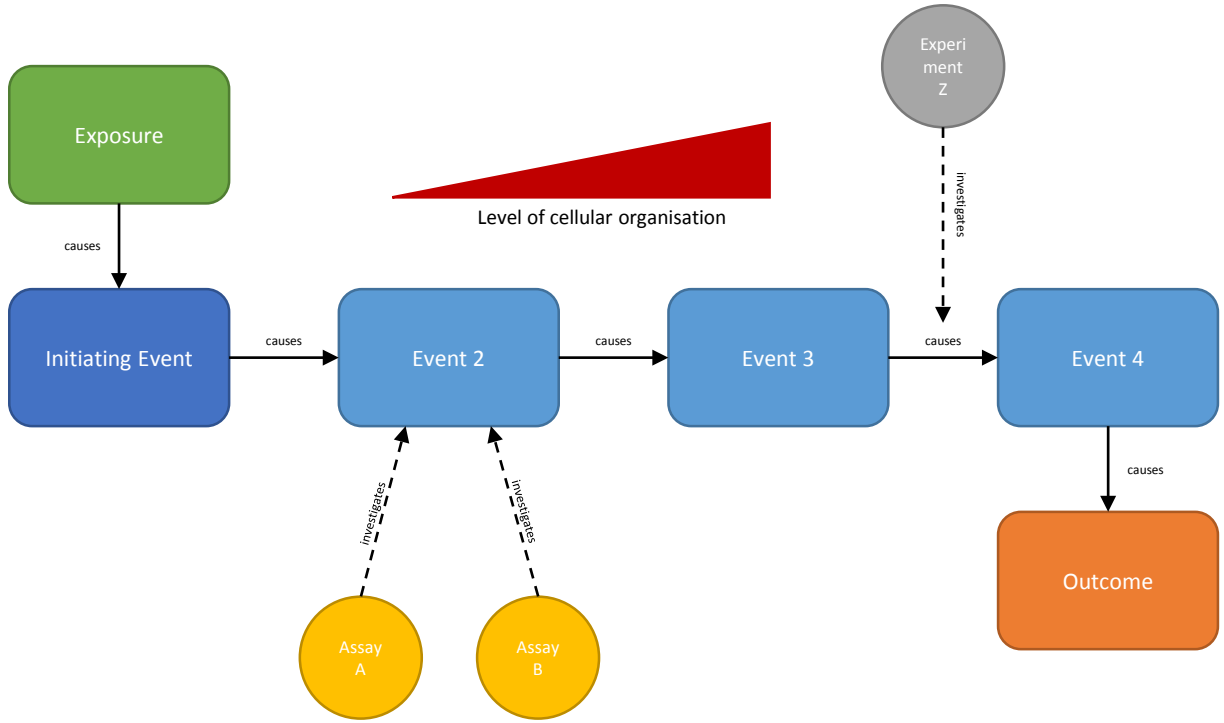
Judgement of similarity at level of

A. whole study
B. broad PECO element
C. individual PECO sub-element

How do we ensure these judgements are valid?

Rules for interpretation? Maybe in AOPs

- Is the observed intermediate event strongly predictive of the target outcome?
- Do the mechanisms in the observed population also happen in the target population?
- Intuition: the more certain the answer, the lower the sense that the evidence is indirect
- If true, maybe judgement of similarity can be derived from a function of certainty in the AOP network
- Potential for grounding judgements of directness in biological knowledge (so long as that knowledge is gathered systematically)



Recap of systematic review and evidence integration
A PECO-based framework for evidence integration
Practical challenges in achieving grounded analysis
Solution: a computational approach
Conclusions and credits

Two major, practical threats to grounded SR

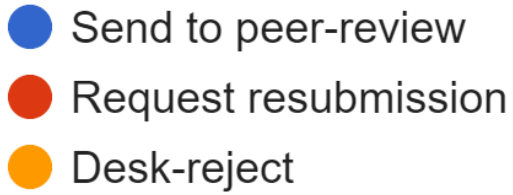
- Implementing valid processes
- Overwhelming data volume

Prepublication data on EH systematic reviews

- At *Environment International*, we triage submissions on six key features of a SR:
  1. Are objectives appropriate to investigating research question?
  2. Does the search methodology miss relevant evidence?
  3. Do the exclusion criteria and screening process exclude relevant evidence?
  4. Have included studies been appraised using a valid risk of bias instrument?
  5. Have appropriate quantitative and qualitative been used to synthesise the evidence?
  6. Has certainty in the evidence been assessed using appropriate, defined criteria?
- We score the methods on a Likert scale of 1-5 (1 = serious concerns)
- A score of 1 or 2 in any domain is a critical shortcoming and results in desk-rejection*

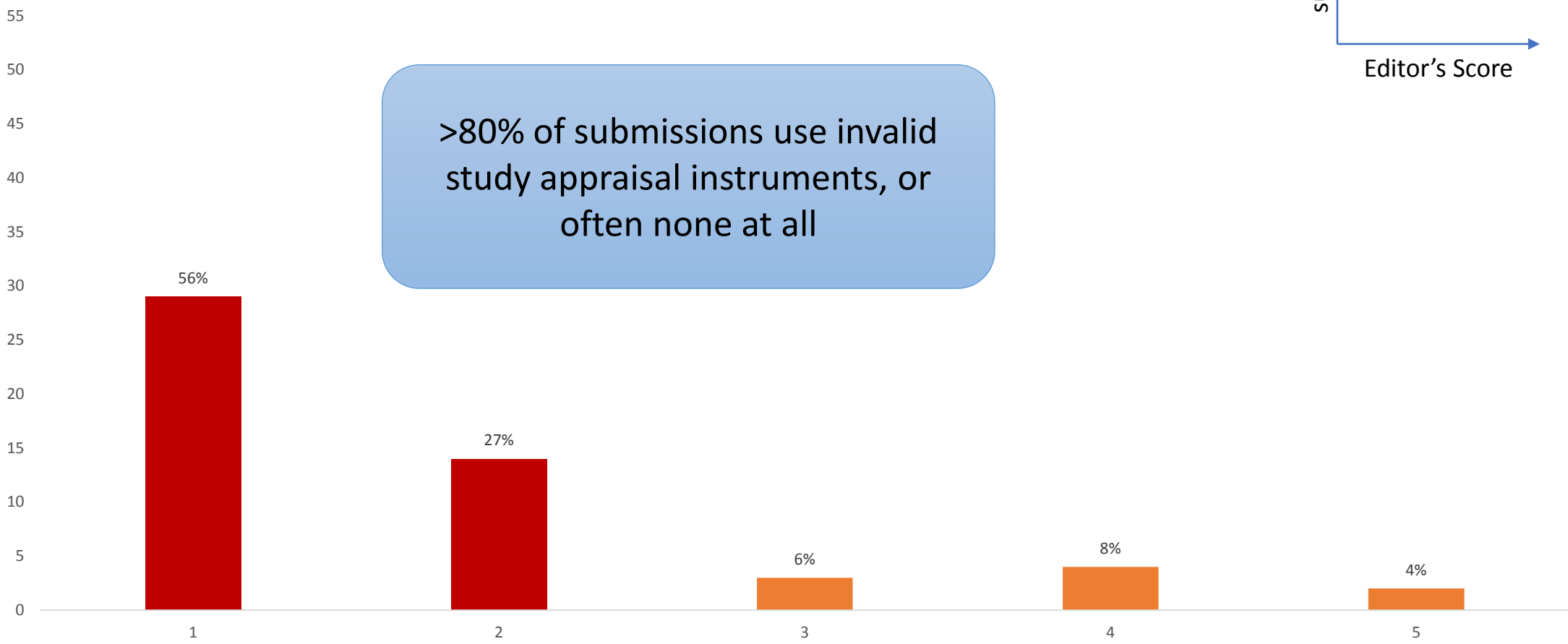
*Authors receive a detailed triage report and editor feedback on identified issues; as often as possible issues are discussed with authors with a view to enabling resubmission

Summary of Triage Decisions

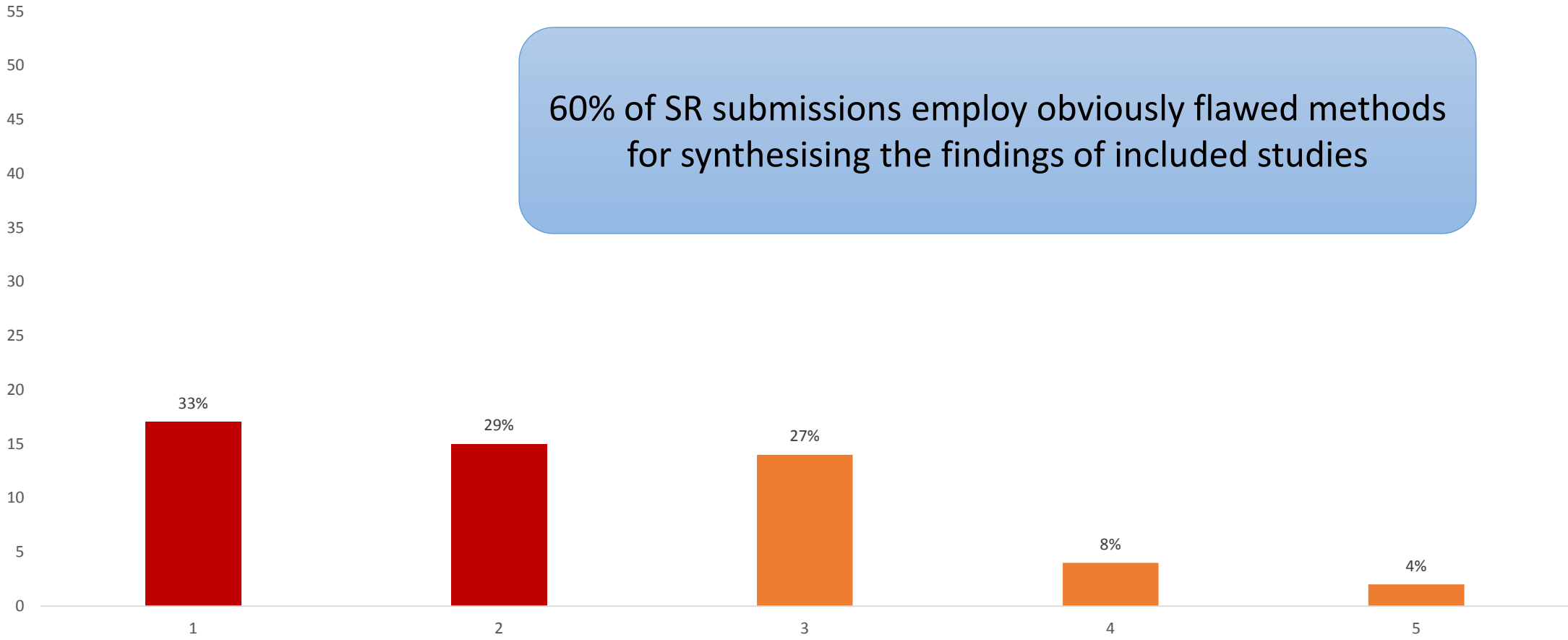


Period April 2018 - May 2019, since introduction of triage tool. n=52

Methods for Study Appraisal

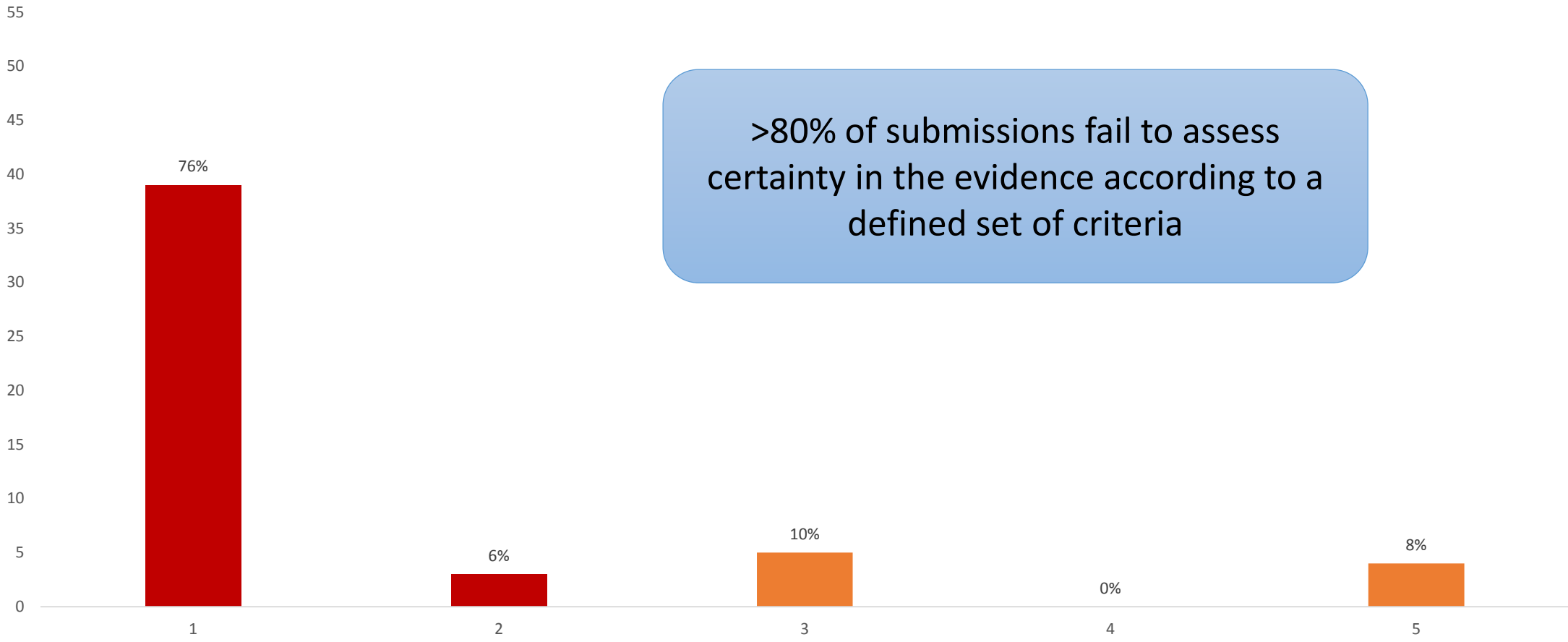


Synthesis methods



60% of SR submissions employ obviously flawed methods for synthesising the findings of included studies

Methods for Certainty Assessment



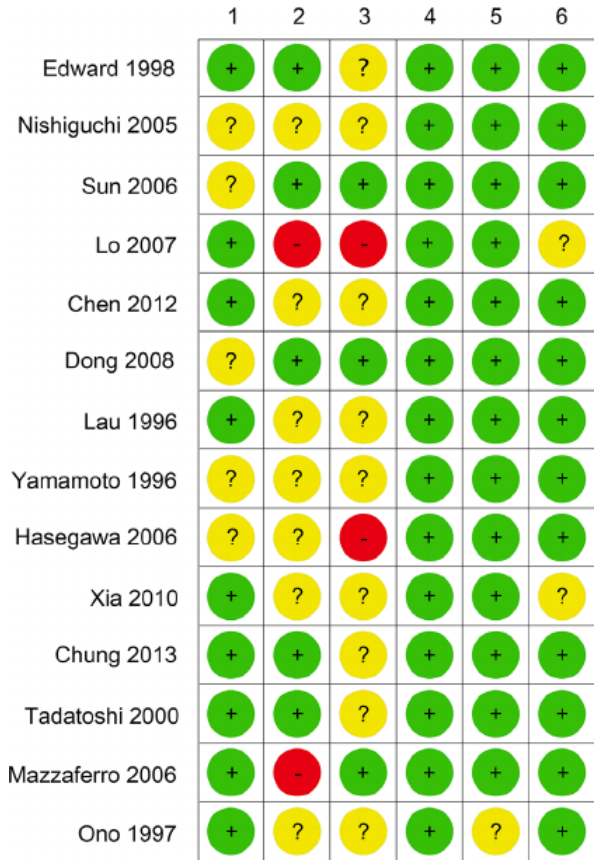
>80% of submissions fail to assess certainty in the evidence according to a defined set of criteria

Post-publication data from medical SRs

- 8989 PubMed records tagged by 2004 as “systematic review” yet actual number of stringently-defined SRs was ≈2500 (Moher et al. 2007)
- Most published SRs have major flaws in conduct and reporting (Page et al. 2016)
- ≈3% of manuscripts are “decent and clinically useful” (Ioannidis 2016)
- What about Cochrane?
  - Propadalo et al. 2019: 29% of Cochrane reviews are discrepant with guidance on allocation concealment
  - Babic et al. 2019 : “Assessments of attrition bias in Cochrane systematic reviews are highly inconsistent”
- These are intervention reviews, not aetiology

Educating our way out of this challenge?

- Most EH research teams do not successfully apply even the simpler, well-documented instruments (e.g. OHAT, Navigation Guide, GRADE) which would better ground their SR methods
- Even if we ended up doing as well on average as the medics, we wouldn't be doing well enough
- Doing as well as the outlier (setting up a Cochrane for EH research) is not a near-future event
- Complex tools like ROBINS-E: what prospects for successful use given the above?

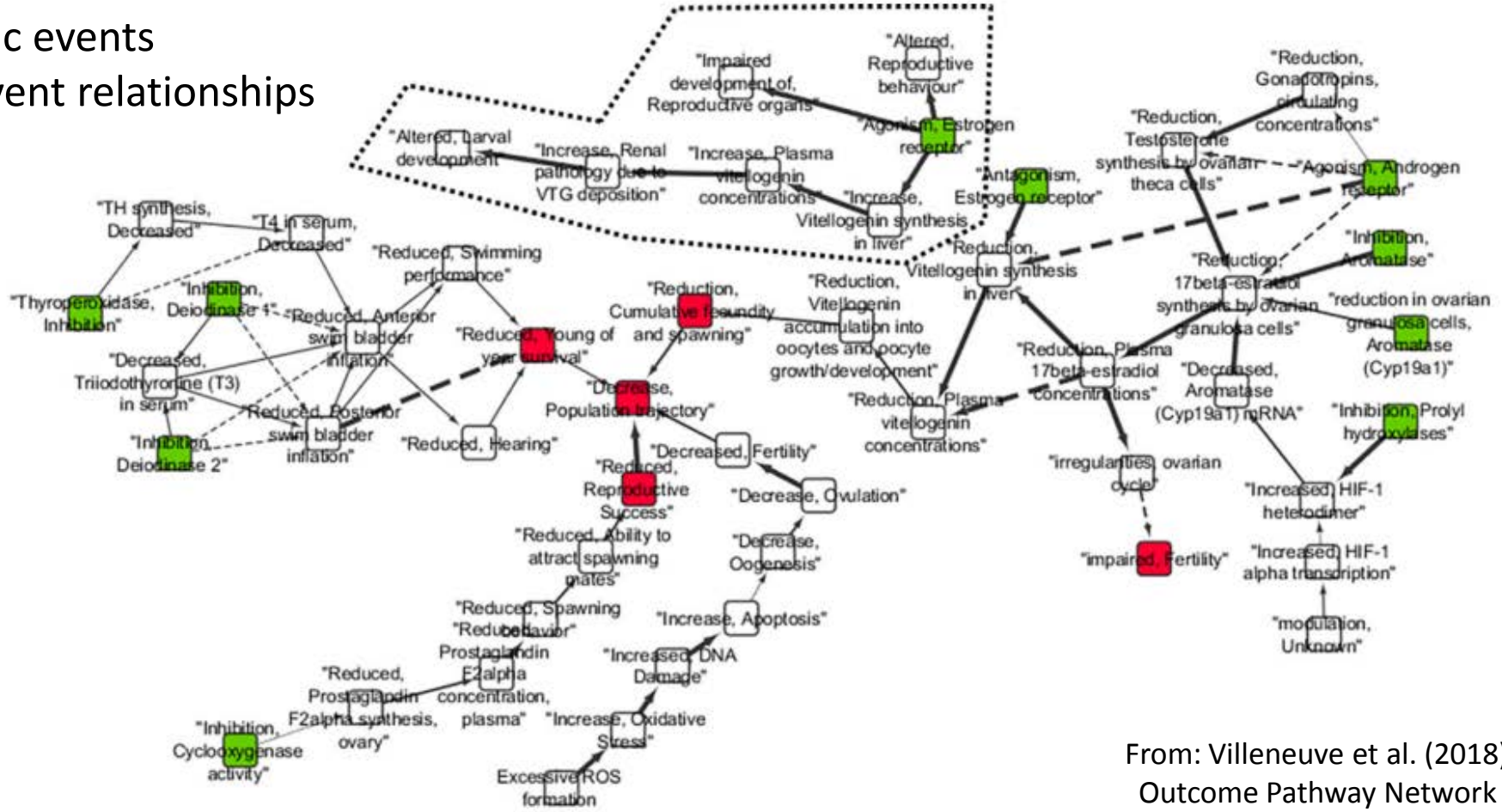


The data volume problem

CYP19 AOP network

>50 biokinetic events

>65 event/event relationships



From: Villeneuve et al. (2018) Adverse Outcome Pathway Network Analytics

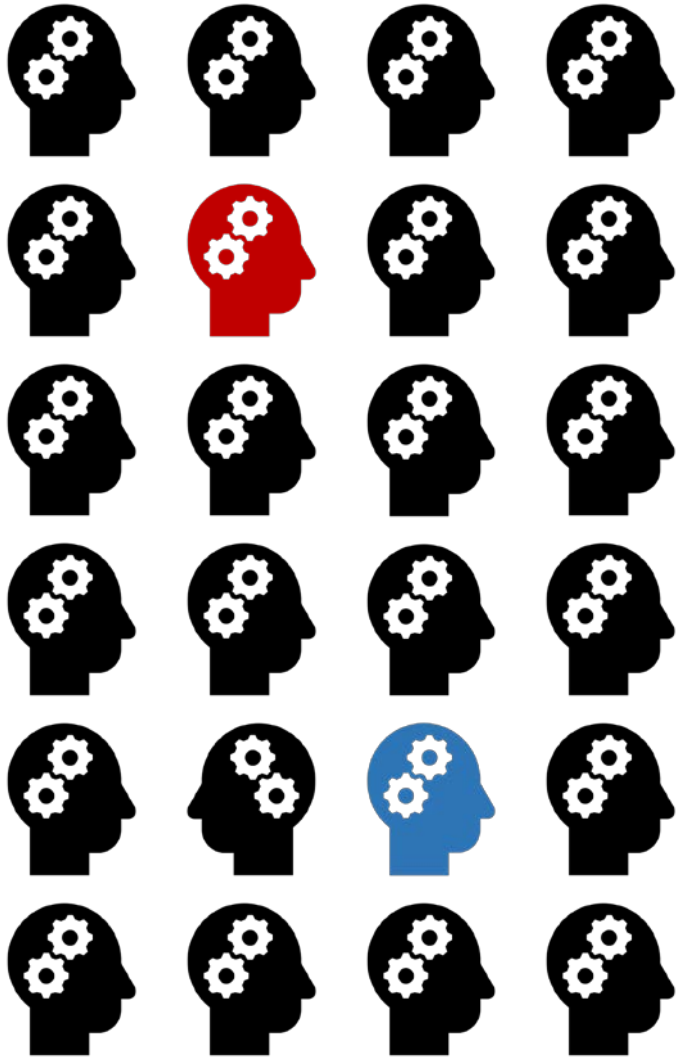
The integration challenge, in a nutshell

- If the stars align, simple SRs can successfully be conducted
- But in most normal scenarios, SR methods are out of reach of most researchers' capacity to apply them successfully
- Methods for integrating mechanistic data into SRs are unlikely to be any easier to apply successfully – plus, they overwhelm us with data
- We can't escape this challenge: the methods need to be applied in order for SRs to be grounded
- So we need a scalable approach to grounded integration methods

Recap of systematic review and evidence integration
A PECO-based framework for evidence integration
Practical challenges in achieving grounded analysis
Solution: a computational approach
Conclusions and credits

In favour of algorithms

- By turning features into numbers, we can make processes repeatable and scalable (i.e. computers can do the work for us)
- Discussable inputs which can be changed deliberately
- The challenge is preserving the links in the chain of evidence that keeps the process grounded (score-text-design-validity)
- How do we do that for complex SR questions, e.g. predicting dose-response relationships in human populations using indirect evidence?



	P features				E features			C features	O features
Study	**Specie**	**L. Org.**	**Age**	**Sex**	**Chem**	**Dose**	**Timing**	**Dose**	**Outcome**
Target	Human	Whole organism	Pre-menopause	Female	OC	1 µg/kg bw	Pre-puberty	1 µg/kg bw increments	Endometriosis
Ref013	Human	Whole organism	Adult	Female	Furan mix	High exposure group	Up to 16 years age	Low exposure group	Endometriosis
Ref852	Human	HESC cells	-	Female	TCDD	10uM solution	-	10 uM increments	Migration
Ref134	Wistar Rat	Whole organism	24 months	Male	Chlorpyrifos	1000 µg/kg bw/d	Until weaning	Vehicle	PR-B/A expression

We can readily turn judgements of similarity into numbers within our tPECO framework

Study	Specie	L. Org.	Age	Sex	Chem	Dose	Timing	Dose	Outcome
Target	Human	Whole organism	Pre-menopause	Female	OC	1 µg/kg bw	Pre-puberty	1 µg/kg bw increments	Endometriosis
Ref013	1	1	2	1	3	4	4	2	1
Ref852	1		-		1	2	-	2	3
Ref134	3	1	2	4	1	6	2	1	4

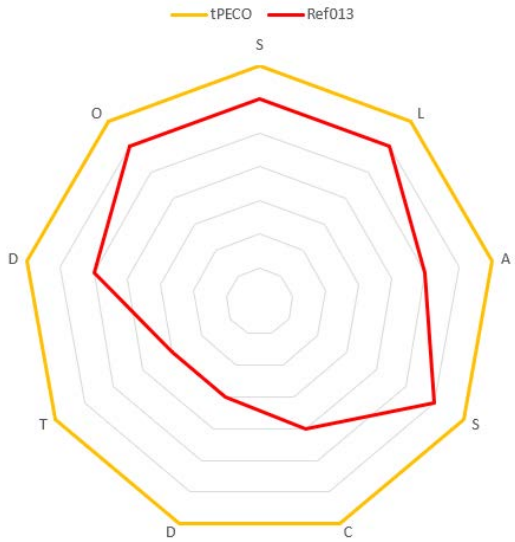
Human

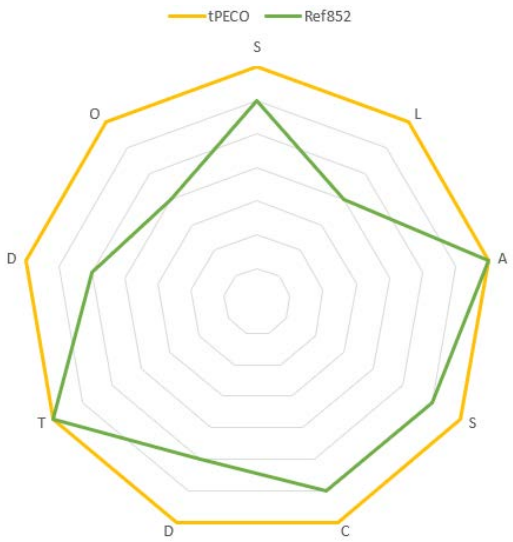
Any adult

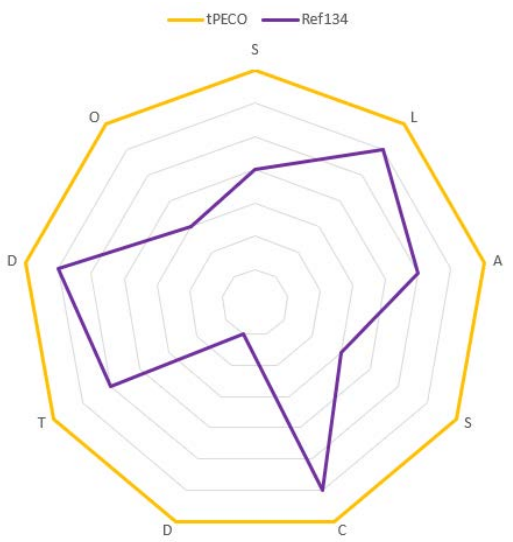
Furans

Up to 16

How similar?







How do we ground similarity scores?

- In our mechanistic study, what makes a rat score a 3? Or PR-B/A a 4?
- The million (multi-trillion?) dollar question

	P features				E features			C features	O features
Study	**Specie**	**L. Org.**	**Age**	**Sex**	**Chem**	**Dose**	**Timing**	**Dose**	**Outcome**
Target	Human	Whole organism	Pre-menopause	Female	OC	1 µg/kg bw	Pre-puberty	1 µg/kg bw increments	Endometriosis
Ref013	1	1	2	1	3	4	4	2	1
Ref852	1	3	-	1	1	2	-	2	3
Ref134	3	1	2	4	1	6	2	1	4

Rat

PR-B/A expression

Research for grounding similarity scores

- Grounding requires us to connect the numbers to the textual record, and to the empirical evidence for their interpretation (their value)
- There are at least three big jobs that need to be done
  1. Systematic methods for AOP development
  2. Automated data extraction
  3. Machine-learning models for weighting evidence
- Probably all three need doing, because it looks like a big-data challenge

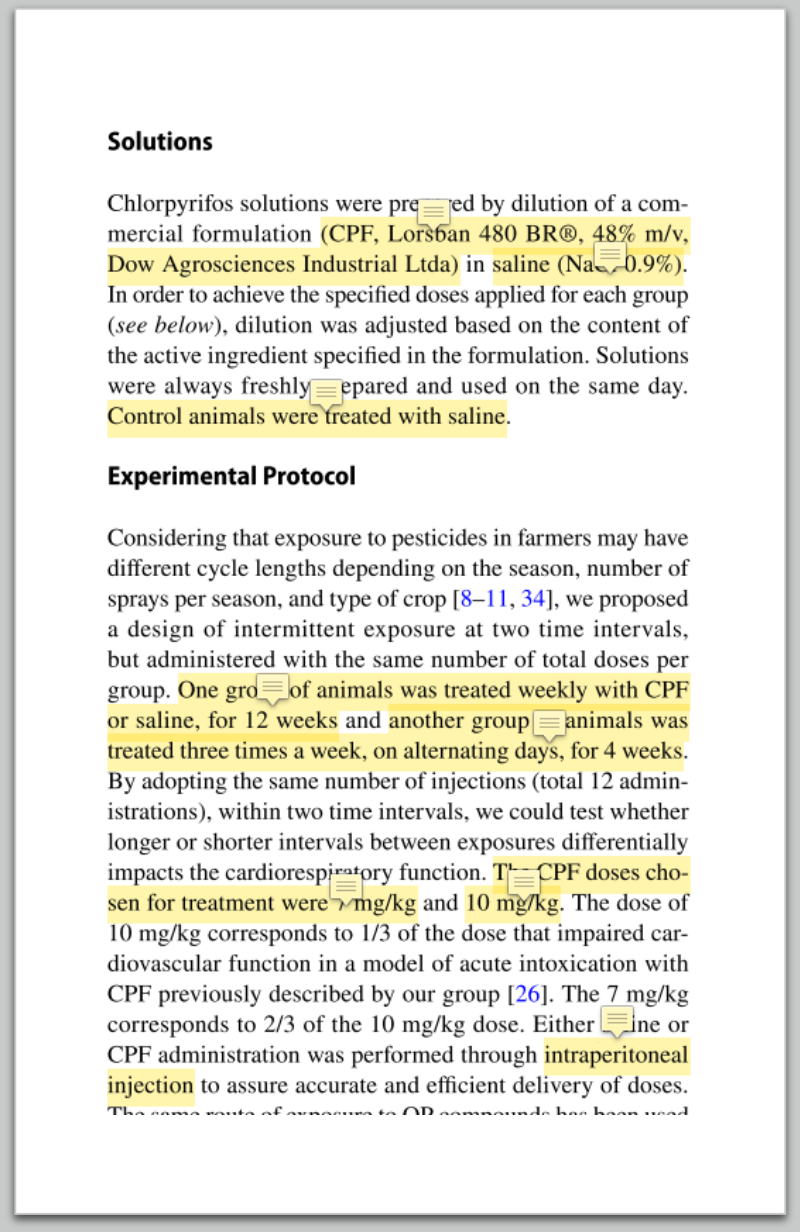
1. Systematic approach to AOP development

- Data model for external validity is underpinned by AOPs
- But we haven't formalised the key features from which AOPs are built
  - What information in the textual record should we use when developing an AOP?
  - What rules should we follow in developing valid AOPs / determining their plausibility?
- This will need to be grounded, and therefore systematic*
- If we figure this out, we will know what rules the machines should be following when identifying and evaluating putative AOPs for us

*SR approach to AOPs is subject of EBTC GRADE pre-meeting in Hamilton next week

2. Automated data extraction

- PECO features and AOP information need extracting from narrative text in full study reports
- This will be a very large extraction job: high level of granularity across thousands of documents
- Would require automation to be practically doable, therefore natural language processing (NLP) approach
- NLP methods can't yet differentiate the features we are interested in, at level of full text, with enough reliability to do data extraction for us
- The step-change which is required implies need for a full-text toxicology corpus training set



Solutions

Chlorpyrifos solutions were prepared by dilution of a commercial formulation (CPF, Lorsban 480 BR®, 48% m/v, Dow Agrosciences Industrial Ltda) in saline (NaCl 0.9%). In order to achieve the specified doses applied for each group (*see below*), dilution was adjusted based on the content of the active ingredient specified in the formulation. Solutions were always freshly prepared and used on the same day. Control animals were treated with saline.

Experimental Protocol

Considering that exposure to pesticides in farmers may have different cycle lengths depending on the season, number of sprays per season, and type of crop [8–11, 34], we proposed a design of intermittent exposure at two time intervals, but administered with the same number of total doses per group. One group of animals was treated weekly with CPF or saline, for 12 weeks and another group animals was treated three times a week, on alternating days, for 4 weeks. By adopting the same number of injections (total 12 administrations), within two time intervals, we could test whether longer or shorter intervals between exposures differentially impacts the cardiorespiratory function. The CPF doses chosen for treatment were 7 mg/kg and 10 mg/kg. The dose of 10 mg/kg corresponds to 1/3 of the dose that impaired cardiovascular function in a model of acute intoxication with CPF previously described by our group [26]. The 7 mg/kg corresponds to 2/3 of the 10 mg/kg dose. Either saline or CPF administration was performed through intraperitoneal injection to assure accurate and efficient delivery of doses.

Rats have four legs, big ears and a tail.

We breed Han-Wistars.

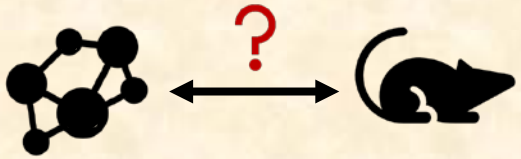
Paul is being ratty because he is tired and hungry.

Paul isn't ratty enough for Warfarin to poison him.

Artists like a golden ratio.

No rats were harmed during filming.

Tree-rats keep stealing food from our bird-feeders.



Teaching computers to read

- Computers "read" by building statistical models to attempt to discern the same regularities in a written document that people respond to when discerning the meaning therein (the written concept "rat" will have a certain statistical shape in a document)
- The problem is there are lots of things which will, to the statistical models, look like regularities which are not meaningful (i.e. look like rats but are not rats), while many meaningful regularities will be invisible to them (are rats, but do not look like them)
- To help, we can manually annotate a large, representative set of documents (a corpus) to show the machines the parts which are meaningful to us (where the rats actually are). The machine can heavily weight this information in its statistical model, massively improving its performance for a data extraction task

Machine-learning models for weighting evidence

- Starts off with responding to the features we know are important (blinding, species, vehicle, event, dose regimen, formulation etc.)
- Uses statistical models of those features to repeat human processes at high volume (e.g. judges risk of bias, indirectness, etc.)
- Large datasets yielded by success with NLP implies quantitative models for interpreting meaning of dataset features
- Over time, the machine identifies predictive features we are not aware of, and improves its performance beyond human capability

Recap of systematic review and evidence integration
A PECO-based framework for evidence integration
Practical challenges in achieving grounded analysis
Solution: a computational approach
Conclusions and credits

Summary

- Successful evidence integration requires us to ground complex judgements of the directness of evidence in (a) the textual record of research and (b) in biological knowledge
- We have proposed a framework for using PECO statements to structure judgments about external validity, which seems to necessitate a computational implementation
- We have outlined a research roadmap toward how such an implementation can be realised and grounded

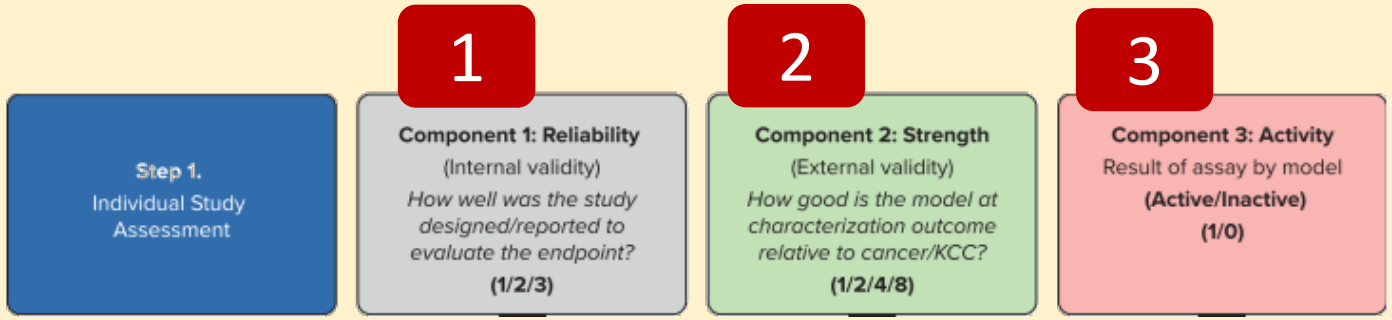
Thanks to…*

- **Stephen Wattam**, WAP Academic Consultancy Ltd
- **Daniele Wikoff**, Toxstrategies LLC
- **Oliver Wild**, Lancaster Environment Centre, UK
- **Taylor Wolffe**, Lancaster Environment Centre, UK
- **Paul Rayson**, Lancaster University School of Computing and Communications
- **John Vidler**, Lancaster University School of Computing and Communications
- **EBTC staff**: Katya Tsaioun, Sebastian Hoffmann, Rob de Vries
- **Patient listeners**: Rebecca Morgan, Michelle Angrish

*Credit is theirs, mistakes are mine

Wikoff Model* for quantitative integration

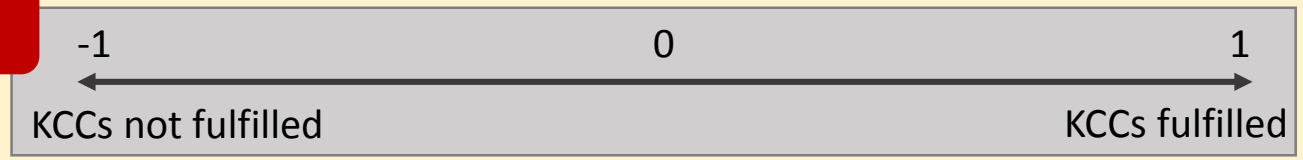
- Model measures the extent to which a body of evidence relevant to the potential carcinogenicity of a chemical fulfils the KCCs
- Uses three inputs (1-3) and an algorithm (4) to provide a numeric description (5) of how well the evidence "matches" the KCCs
- It works a bit like calculating Flesch-Kincaid readability scores in word processors: overall target characteristic described as a function of some measurable properties, normalised onto a scale



4

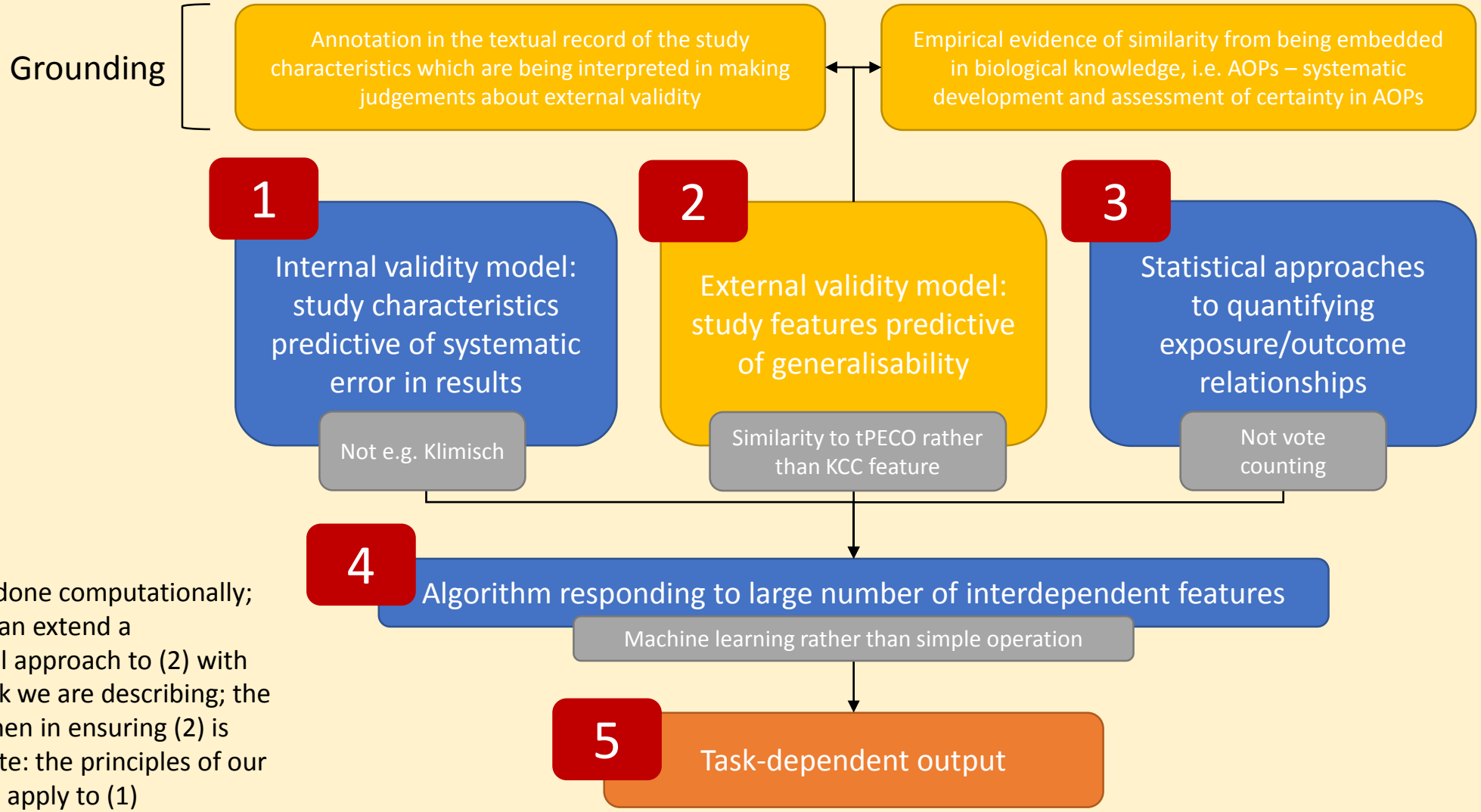
$$\frac{\sum_{i,\mathrm{all}}\left[\left(\frac{w_R(-R_i+4)+w_M M_i}{w_R R_{\max}+w_M M_{\max}}\right)*\frac{E_{\mathrm{Act},i}}{E_{\mathrm{Total},i}}-\left(\frac{w_R(-R_i+4)+w_M M_i}{w_R R_{\max}+w_M M_{\max}}\right)*\frac{E_{\mathrm{InAct},i}}{E_{\mathrm{Total},i}}\right]}{n_{i,\mathrm{all}}}$$

5



*Oversimplified version presented here, see Wikoff et al. (2019) for detail

From Wikoff to grounded integration



(3) is already done computationally; we think we can extend a computational approach to (2) with the framework we are describing; the challenge is then in ensuring (2) is grounded. Note: the principles of our approach also apply to (1)