Artificial Intelligence and Systematic Review

Paul Whaley, PhD

Co-Chair, Open Science Working Group

Evidence-Based Toxicology Collaboration

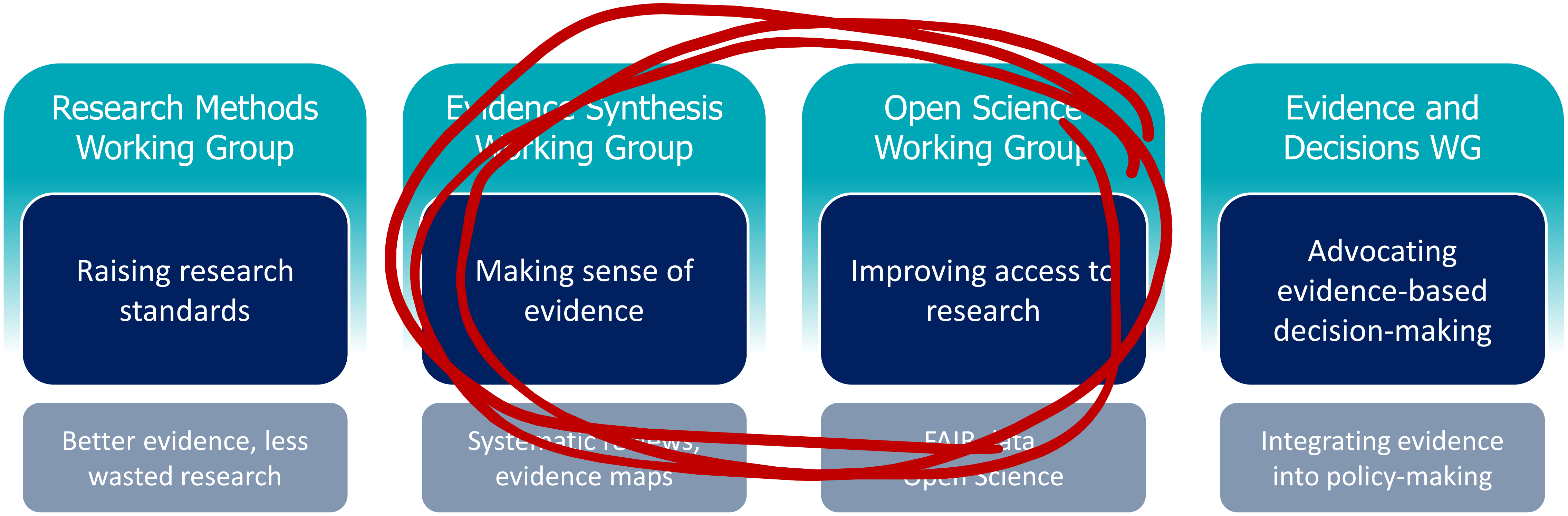
EFSA Advisory Forum, 1 December 2023

What is EBTC?

An international collaboration improving how we create, use, and publish toxicological and environmental health research

Raising research standards	Making sense of evidence	Improving access to research	Advocating evidence-based decision-making
Better evidence, less wasted research	Systematic reviews, evidence maps	FAIR data, Open Science	Integrating evidence into policy-making

A working group for each “pillar”



Research Methods Working Group

Raising research standards

Better evidence, less wasted research

Evidence Synthesis Working Group

Making sense of evidence

Systematic reviews, evidence maps

Open Science Working Group

Improving access to research

FAIR data, Open Science

Evidence and Decisions WG

Advocating evidence-based decision-making

Integrating evidence into policy-making

EBTC work in evidence synthesis and AI

- Using automation tools in very large systematic reviews (screening 15-30k papers)
- Proof-of-concept for machine-supported approaches to improving conduct and reporting of research
- Co-leading GRADE Ontology project, for consistent machine-readable expression of concepts of certainty of evidence and strength of recommendations
- Established *Evidence-Based Toxicology* journal for innovation in open science publishing

EVIDENCE-BASED TOXICOLOGY

The Official Journal of the Evidence-Based Toxicology Collaboration

VOLUME 1 ISSUE 1 2023

EVIDENCE-BASED TOXICOLOGY
2023, VOL. 1, NO. 1, 2259938
https://doi.org/10.1080/2833373X.2023.2259938

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

RESEARCH ARTICLE

Using bench protocol platforms to improve compliance with systematic review guidance documents and reporting checklists: Proof of concept

Paul Whaley[a], Stephen Wattam[b], Anna Mae Scott[c], John Vidler[d] and Evidence-Based Toxicology Collaboration[e]

[a]WhaleyResearch and Lancaster Environment Centre, Lancaster University, Lancaster, UK; [b]WAP Consulting Limited, Northallerton, UK; [c]Institute for Evidence-Based Healthcare, Bond University, Gold Coast, Australia; [d]School of Computing and Communications, Lancaster Univ... UK; [e]Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

...checklists have been ...rigorous

ARTICLE HISTORY
Received 12 January 2023
Revised 28 August 2023
Accepted 12 September

scientific reports

www.nature.com/scientificreports

OPEN

Check for updates

Performance of preclinical models in predicting drug-induced liver injury in humans: a systematic review

Hubert Dirven[1], Gunn E. Vist[2], Sricharan Bandhakavi[3], Jyotsna Mehta[4], Seneca E. Fitch[5], Pandora Pound[6], Rebecca Ram[6], Breanne Kincaid[7], Cathalijn H. C. Leenaars[8], Minjun Chen[9], Robert A. Wright[10] & Katya Tsaioun[11]

Drug-induced liver injury (DILI) causes one in three market withdrawals due to adverse drug reactions, causing preventable human suffering and massive financial loss. We applied evidence-based methods to investigate the role of preclinical studies in predicting human DILI using two anti-diabetic drugs from the same class, but with different toxicological profiles: troglitazone (withdrawn from US market due to DILI) and rosiglitazone (remains on US market). Evidence Stream 1: A systematic literature review of in vivo studies on rosiglitazone or troglitazone was conducted (PROSPERO registration CRD42018112353). Evidence Stream 2: in vitro data on troglitazone and rosiglitazone were retrieved from the US EPA ToxCast database. Evidence Stream 3: troglitazone- and rosiglitazone-related

Who am I?

- Co-chair of EBTC Open Science WG
- Editor-in-Chief, *Evidence-Based Toxicology*. Formerly SRs editor at Environment International (450 SRs edited since 2016)
- Seen first-hand how SR defines data requirements and processes for making sense of evidence, while the demands on people to do SR combined with explosion in publication of research means it is impossible to feed society’s knowledge needs without AI being involved

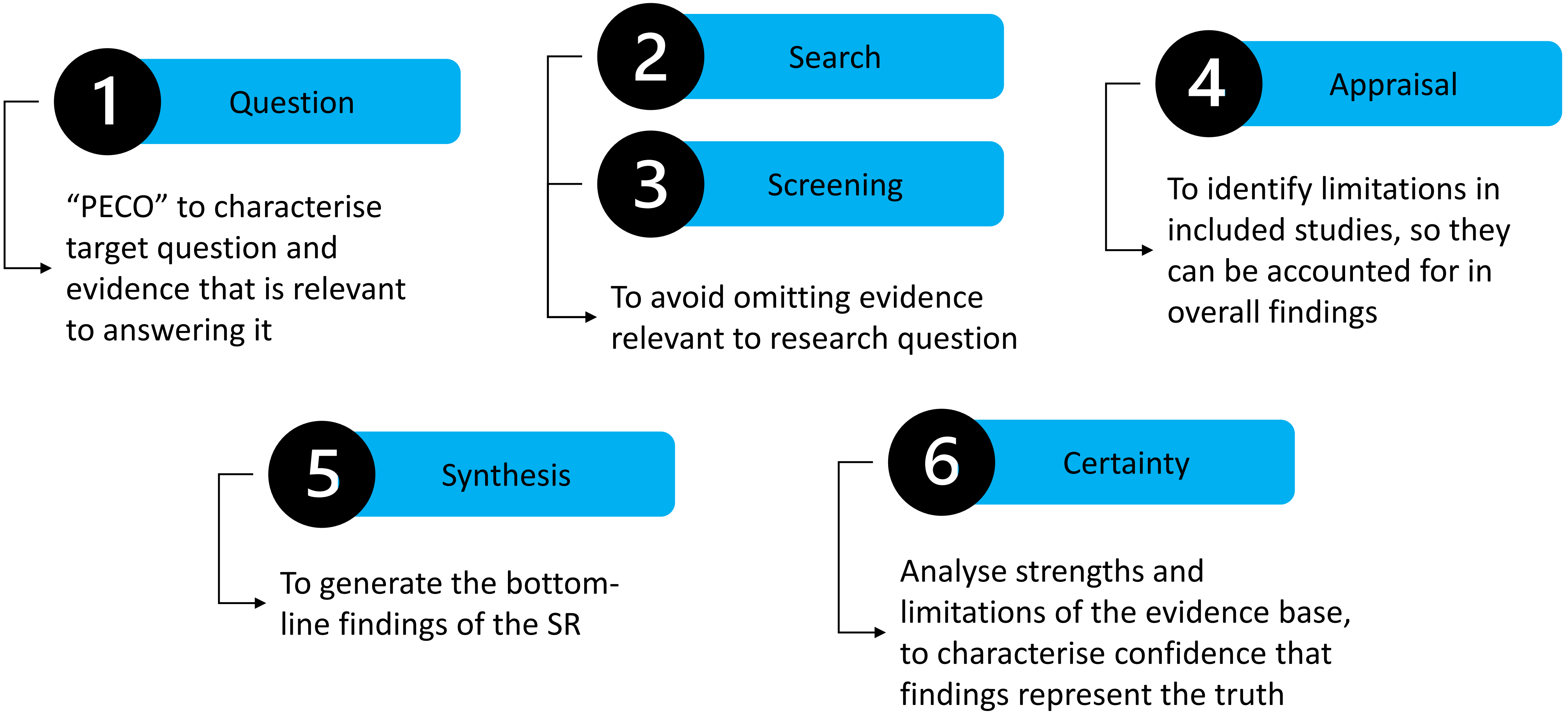
Why do we need Artificial Intelligence in Systematic Review?

| *Because we need SR, but it is too slow* |

SR as a fundamental of EBT

- Apply scientific method when using existing evidence to answer research questions
- At each stage of review process, minimise potential for bias and random error, maximise transparency and reproducibility

Components of systematic review



1 Question

"PECO" to characterise target question and evidence that is relevant to answering it

2 Search

3 Screening

To avoid omitting evidence relevant to research question

4 Appraisal

To identify limitations in included studies, so they can be accounted for in overall findings

5 Synthesis

To generate the bottom-line findings of the SR

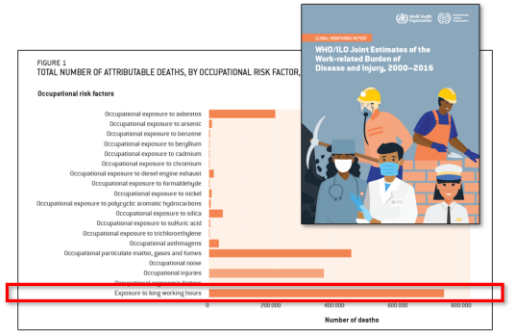
6 Certainty

Analyse strengths and limitations of the evidence base, to characterise confidence that findings represent the truth

SR in toxicology and env health

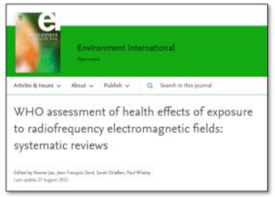
WHO/ILO Burden of Disease

- Ground-breaking in use of protocols and rigour of SR methods
- Confidence to conclude that long working hours is largest cause of workplace mortality (bigger than injury and air pollution)



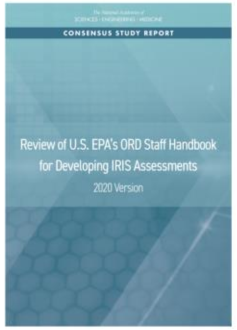
WHO EMF systematic reviews

- Applying same approach from WHO/ILO BoD SRs to evaluating health effects of EMF exposure

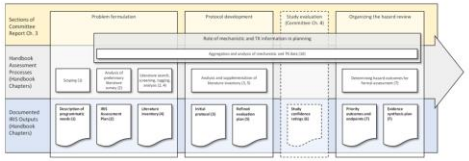


WHO assessment of health effects of exposure to radiofrequency electromagnetic fields: systematic reviews

NASEM review of IRIS Handbook

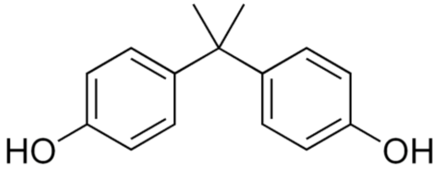

The National Academies of SCIENCES · ENGINEERING · MEDICINE
CONSENSUS STUDY REPORT
Review of U.S. EPA's ORD Staff Handbook for Developing IRIS Assessments
2020 Version

The committee found that the handbook reflects the significant improvements that EPA has made in its IRIS assessment process. For instance, the handbook describes the inclusion of sophisticated, state-of-the-art methods that use systematic evidence maps to summarize literature characteristics for scoping and systematic review methods for hazard identification. Moreover, the IRIS program is clearly helping to advance the science of systematic review, as applied to hazard identification. EPA staff are actively involved in the ongoing development of methods, such as study evaluation and handling of mechanistic data. The committee recognizes that EPA faces challenges in implementing many of the methods for the IRIS assessment process and is impressed and encouraged by the


Sections of Committee Report Ch. 3
Problem formulation
Protocol development
Study evaluation (Committee Ch. 4)
Organizing the hazard review
Role of mechanistic and TK information in planning
Handbook Assessment Processes (Handbook Chapters)
Documented IRIS Outputs (Handbook Chapters)

EFSA 2023 re-evaluation of BPA

- 2015 assessment (not systematic): TDI of 4 μg/kg bw/d
- 2023 assessment (systematic): TDI of 0.2 ng/kg bw/d



SR is a slow process

- Typically searching and screening of 4,000 to 40,000 references
- Detailed analysis of 15 to 150 studies
- Approximately 18 months for a small SR
  - 4,000-10,000 references screened, 15-40 studies analysed in detail
- About 3 years to do a very large SR
  - 25,000-40,000 references screened, 100-150 studies analysed in detail

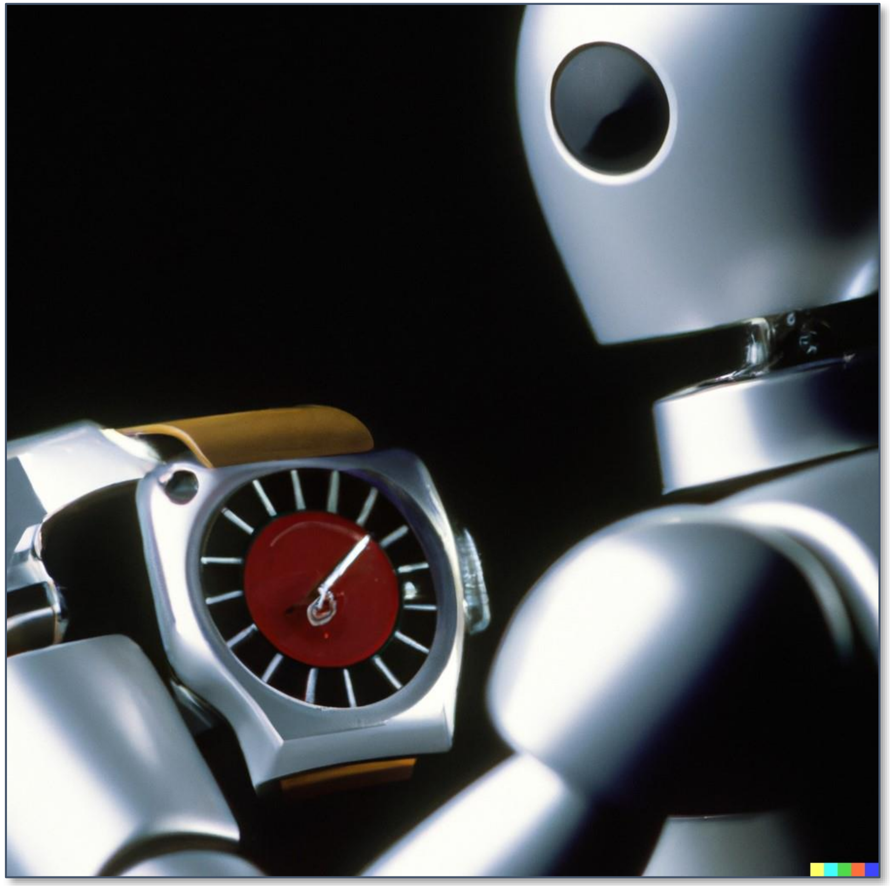
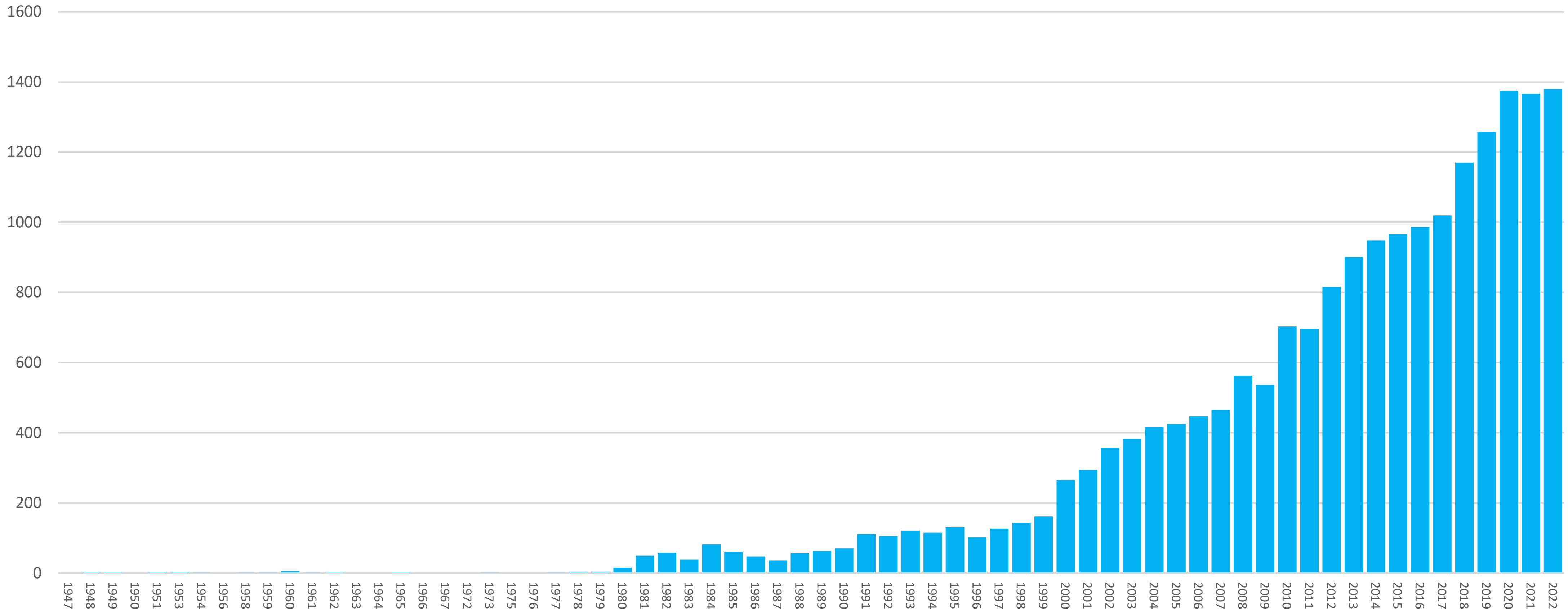


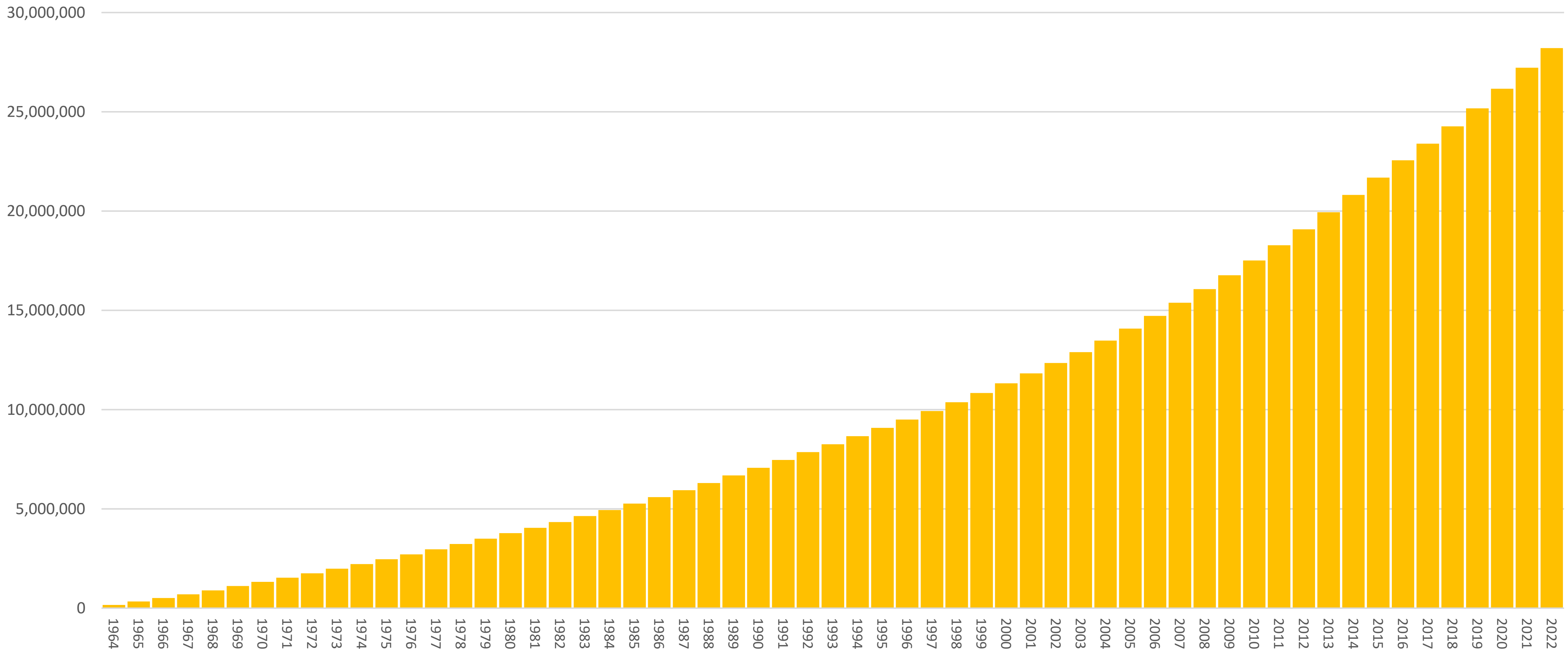
Image created with the assistance of DALL-E2 (OpenAI, 2023)

Haddaway & Westgate (2018) https://doi.org/10.1111/cobi.13231

19,500 studies of BPA: 1,500/yr.



MEDLINE: 1 million per year in 2021



Why is systematic review so slow?

Because this 👇 is scientific knowledge in its wild state

Images created with the assistance of DALL-E2 (OpenAI, 2023)

Can’t afford for it to be so slow

- Policy makers have lots of urgent questions, now
- Been using computers to accelerate evidence synthesis for decades: no coincidence that systematic review took off in the 1990s, when electronic literature searches became possible
- ML-based support for screening decisions and quality appraisal since around 2005*
- “Artificial intelligence” just the latest tools in this process
- Not much has changed, yet as of Nov 2022 everything is changing: where is AI being used now in SR, and what happens next?

*Wallace et al. (2010) https://doi.org/10.1186/1471-2105-11-55

Where is AI helping with systematic reviews?

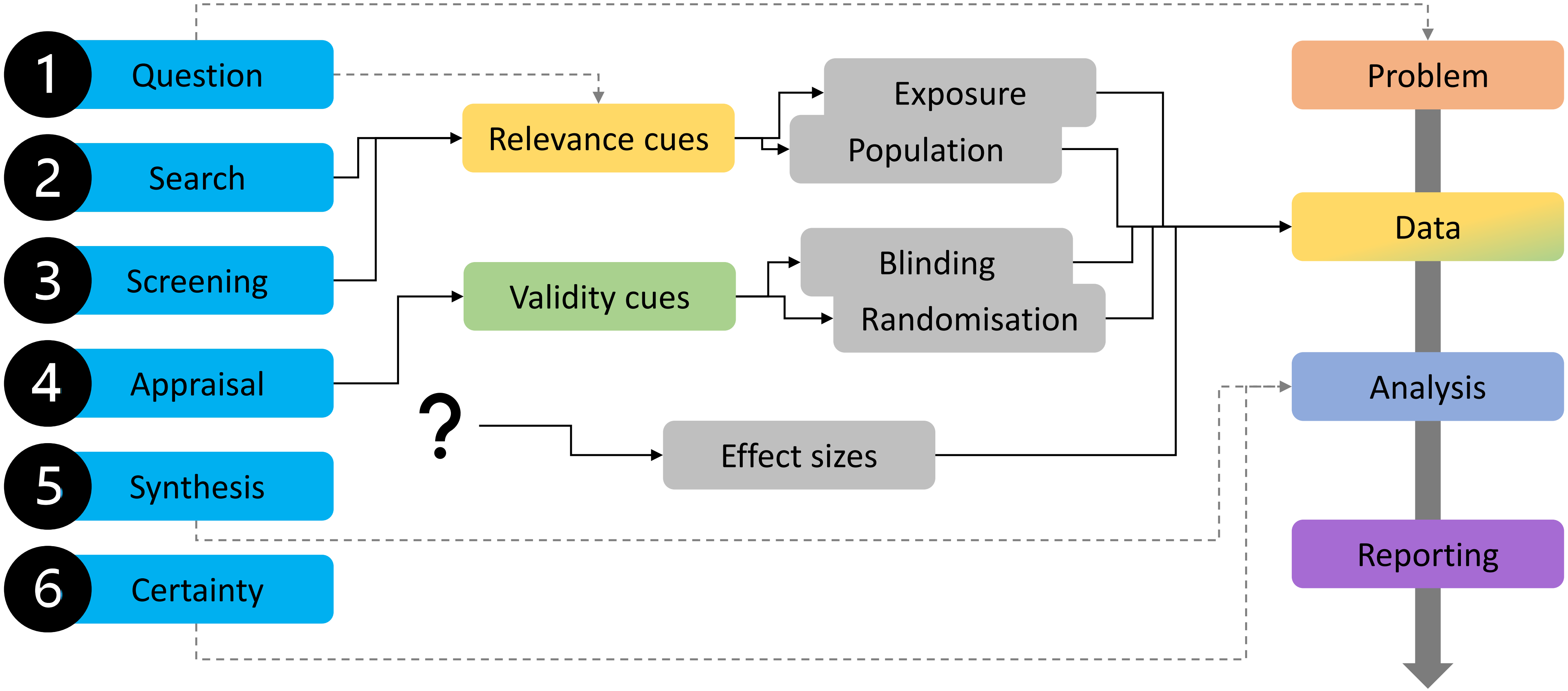
*| EBTC's (*my) experience and perspective: the "old", transitional, and future |*

I had to pause & think about this

1. Question
2. Search
3. Screening
4. Appraisal
5. Synthesis
6. Certainty

- SR traditionally viewed as 6-step, manual process… while AI reframes problems
- Assumes that evidence synthesis is about sifting through individual documents that report scientific studies
- Disaggregates e.g. search and screening into distinct tasks, may be functionally the same
- Neglects tasks of data abstraction and reporting
- Wrong model for a good answer to question of where AI can or should be used in evidence synthesis

From human to machine workflows



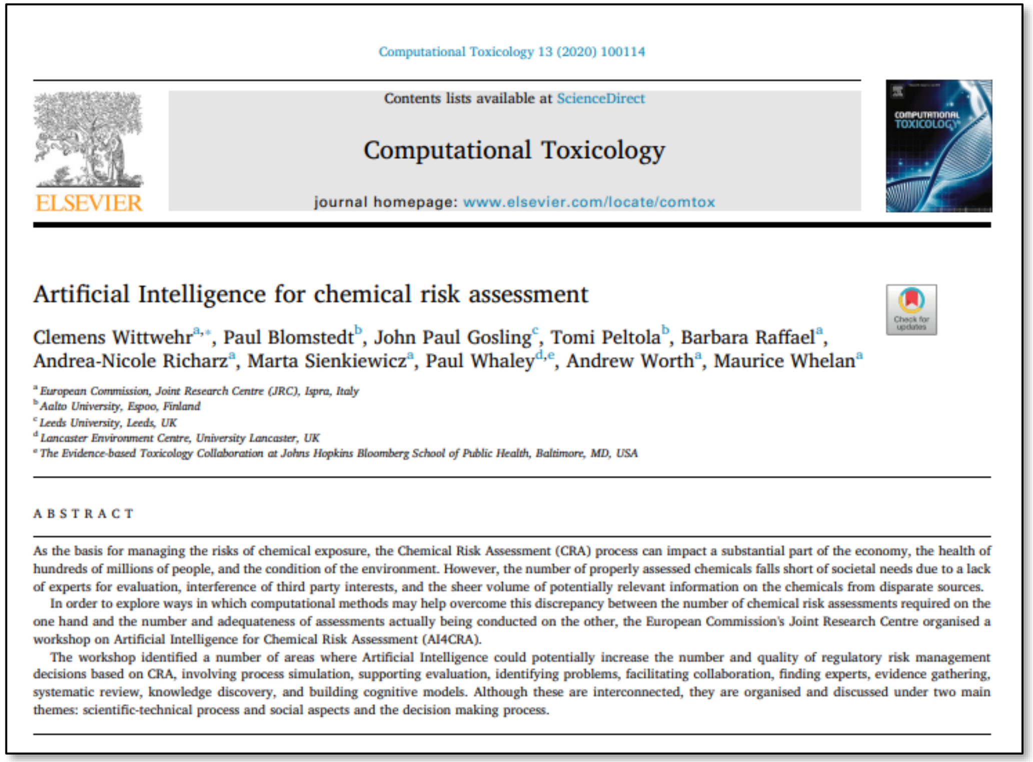
SR in the age of computing / AI

Rethinking the tasks involved in evidence synthesis

a. Define question
b. Retrieve data
c. Analyse data
d. Report findings

a. Defining the question

- **Old way.** Scoping review, asking experts
- **Transition.** Evidence map databases (e.g. US EPA’s HAWC) are replacing scoping reviews. Increasing NLP-based data abstraction (e.g. Dextr)
- **Future?** JRC workshop 2018: blue-skies discussion of what AI might bring*
  - LLMs have made some of these ideas real, e.g. support expert knowledge elicitation
  - Challenges with AI being creative and anticipating policy needs



Computational Toxicology 13 (2020) 100114

Contents lists available at ScienceDirect

Computational Toxicology

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

ELSEVIER

Artificial Intelligence for chemical risk assessment

Clemens Wittwehr[a,*], Paul Blomstedt[b], John Paul Gosling[c], Tomi Peltola[b], Barbara Raffael[a], Andrea-Nicole Richarz[a], Marta Sienkiewicz[a], Paul Whaley[d,e], Andrew Worth[a], Maurice Whelan[a]

[a] *European Commission, Joint Research Centre (JRC), Ispra, Italy*
[b] *Aalto University, Espoo, Finland*
[c] *Leeds University, Leeds, UK*
[d] *Lancaster Environment Centre, University Lancaster, UK*
[e] *The Evidence-based Toxicology Collaboration at Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA*

ABSTRACT

As the basis for managing the risks of chemical exposure, the Chemical Risk Assessment (CRA) process can impact a substantial part of the economy, the health of hundreds of millions of people, and the condition of the environment. However, the number of properly assessed chemicals falls short of societal needs due to a lack of experts for evaluation, interference of third party interests, and the sheer volume of potentially relevant information on the chemicals from disparate sources.

In order to explore ways in which computational methods may help overcome this discrepancy between the number of chemical risk assessments required on the one hand and the number and adequateness of assessments actually being conducted on the other, the European Commission's Joint Research Centre organised a workshop on Artificial Intelligence for Chemical Risk Assessment (AI4CRA).

The workshop identified a number of areas where Artificial Intelligence could potentially increase the number and quality of regulatory risk management decisions based on CRA, involving process simulation, supporting evaluation, identifying problems, facilitating collaboration, finding experts, evidence gathering, systematic review, knowledge discovery, and building cognitive models. Although these are interconnected, they are organised and discussed under two main themes: scientific-technical process and social aspects and the decision making process.

*https://doi.org/10.1016/j.comtox.2019.100114

b. Retrieving data

- **Old way.** Sensitive research index search, then screen (ti/ab easy decisions, full text hard decisions), then abstract data
- Why? Product of severely limited metadata (sometimes only 1% of search results are relevant), inefficiency of reading full texts for the extra information needed, lack of efficient access to data contained in scientific documents
- **Transition.** Human-in-the-loop classifiers emulate reviewer decisions (e.g. SWIFT, Rayyan, etc.); some tools help construct more specific searches. NLP models accelerate abstraction of data from documents identified as relevant (e.g. Dextr)
- **Future?** Machines can rapidly parse full texts and other data. As access to content increases, dependence on human-made metadata decreases. Screening disappears, question becomes how accurate is the data in a database. (OpenAIre? Lens.org? Semantic Scholar?)

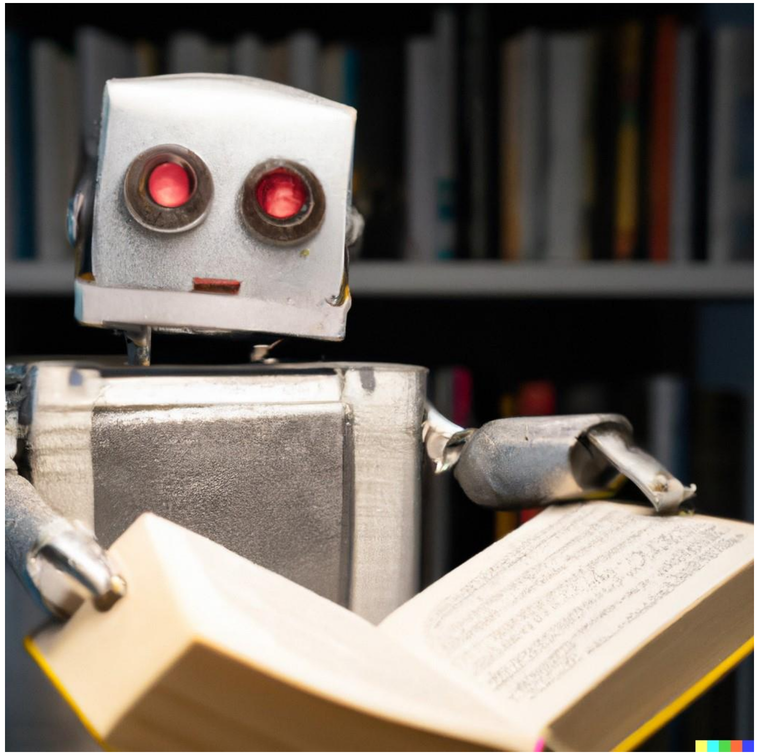


Image created with the assistance of DALL-E2 (OpenAI, 2023)

c. Analysing data

Three components

- Study appraisal: how much weight should we put on units of data that we are analysing?
- Data synthesis: what does the data mean in relation to the question we are answering?
- Certainty assessment: how certain is the answer to the question?

c1. Study appraisal

- **Old way.** Make judgements based on methods data related to potential for bias, broken down by domains (performance, selection, reporting, etc.)

- **Transition.** E.g. RobotReviewer. Is hard for humans, therefore hard for machines
  - Only small number of cues in a document, draws extensively on experience
  - Can still build training sets of expert appraisals – but a big enough training set with enough judgements? How well does it work on studies it has not seen? How often will an appraisal not encounter novel issues with studies?
  - Machine in the loop, like for problem formulation challenges (if the human expert is fundamental to process)

- **Future?** Non-RoB methods? Triangulation surely more accessible using machines. Move away form risk of bias assessment to other methods for characterising potential for error in body of evidence

Image created with the assistance of DALL-E2 (OpenAI, 2023)

c2. Synthesis

- **Old way.** Meta-analysis of studies subjectively judged to be oranges-and-oranges comparisons
- **Future?** ML and computational access to data allows for data models far more sophisticated than the humble meta-analysis (e.g. ONTOX).
- More like integration than synthesis. Will be interesting to see how this pans out

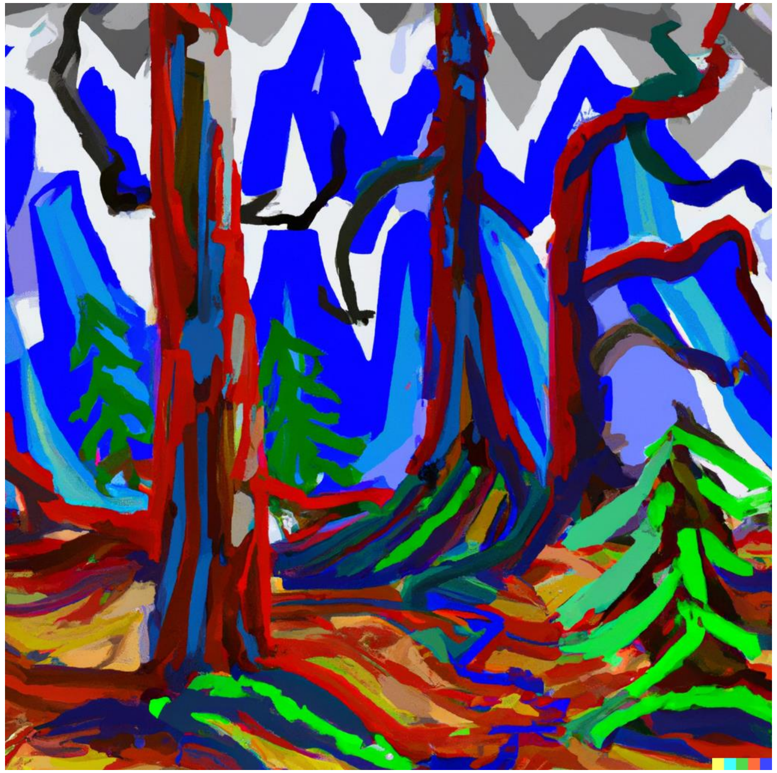


Image created with the assistance of DALL-E2 (OpenAI, 2023)

c3. Certainty assessment

- **Old way.** Expert judgement across certainty domains (e.g. GRADE). A lot of guidance and experience in application needed.
- Similar challenges to risk of bias: highly expert-dependent and interpretive, as much about justification as judgement
- **Future?** Machine in the loop? Training sets, so a machine can emulate a human? Other certainty models?
- Appraisal then synthesis then certainty assessment is probably not the right way to conceive of interpreting a single data set when all parts of the set are computationally accessible

d. Reporting the synthesis

- **Old way.** Inaccurate and incomplete write-up of methods and results, distributed as a PDF by scientific journals
- **Future?** This is where LLMs really come in. Did a proof-of-concept study of automated documentation and standards compliance in summer of 2022 - before LLMs (chat GPT 3.5 November 2022)
- Feed LLMs structured information with a large corpus of SRs to refer to, easy step-change in progress on standards-compliant (PRISMA, COSTER, etc.) first drafts of SR manuscripts
- Applies equally to primary research, and more revolutionary: fills the holes in reporting that make evidence synthesis so challenging – and is more complete data for training the machine models

Whaley et al. (2023) https://doi.org/10.1080/2833373X.2023.2259938

What happens next?

| *The future of evidence synthesis* |

The future is data, all the way down

- Already glimpsed the future: systematic evidence maps - abstract data once, don't look at the documents again

- But if it's data we want, why are we spending so much energy on abstracting data from documents? Because, unfortunately, that is how data is currently recorded and shared

- **Need to break out of the paper-based paradigm**. Scientific publications are a bad way of packaging data. Should be a secondary product of research, with structured labelled data the primary product

- **Systematic review should disappear**: it is a process designed around the limitations of human capacity working in a paper-based research data paradigm. Synthesis should be about manipulating rich research data using models of unfathomable complexity (as glimpsed in ONTOX).

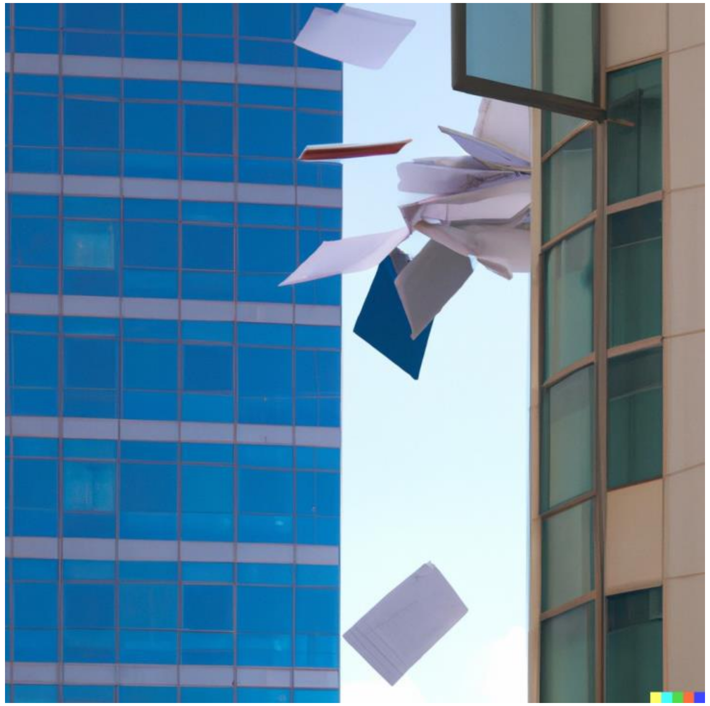


Image created with the assistance of DALL-E2 (OpenAI, 2023)

Wolffe et al. (2020) https://academic.oup.com/toxsci/article/175/1/35/5756220
Whaley et al. (2023) https://doi.org/10.1080/2833373X.2023.2259938

Papers still have a place!

- Obviously not getting rid of scientific papers – it's how humans communicate with each other about research
- Paper just become secondary to data – derived from the data, not the thing from which the data is derived

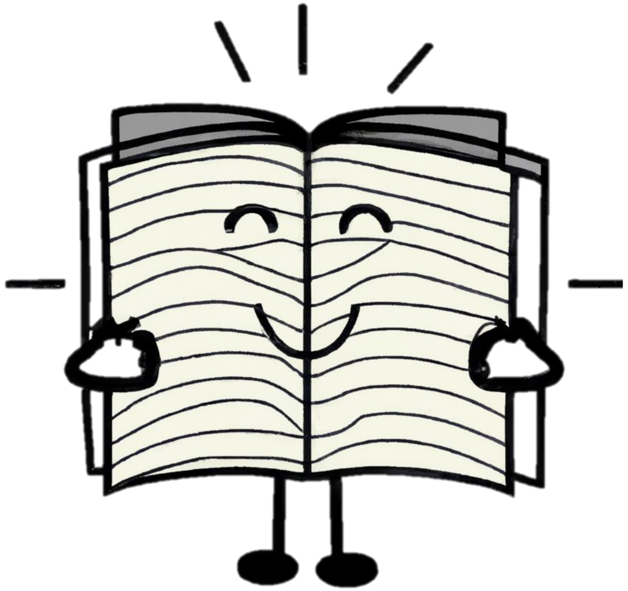


Image created with the assistance of DALL-E2 (OpenAI, 2023)

What the future requires

- Need ontologies (e.g. GRADE Ontology, SEVCO, gene ontology etc. – human made?) and MUCH better metadata (machine-made?) for scientific data to be more accessible to machines
- Machine-readable research (semantic authoring) to help researchers integrate ontologies into reporting of research
- More comprehensive, accurate, reporting of primary studies. Goes hand-in-glove with better SR: cannot analyse data that has not been recorded
- Improved language training sets (from personal, somewhat bitter experience)
- **AI is the exciting part that gets the limelight, but infrastructure which helps computers work better must not be neglected**

A final warning

- Machine processes present a major challenge of understandability - maybe even commensurability
- Doesn't matter how Google's image classifiers or OpenAI's language or image generators work, but generating knowledge is different
- **"If a lion could speak, we could not understand him." – Wittgenstein**
- Maybe the production of coherent, auditable research becomes the test of the system, whether it is AI or human based

Image created with the assistance of DALL-E2 (OpenAI, 2023)

Join EBTC's Open Science WG

- There is also an evidence synthesis WG for those of you into SR
- Members support each other in conducting high-quality, high-impact projects of strategic value to the EBT community
- To join the collaboration or propose project ideas, email me or click here!

paul@whaleyresearch.uk