frontiers in Pharmacology

HYPOTHESIS AND THEORY
published: 17 December 2019
doi: 10.3389/fphar.2019.01317

E-Synthesis: A Bayesian Framework for Causal Assessment in Pharmacosurveillance

Francesco De Pretis[1,2†], *Jürgen Landes*[3†] *and Barbara Osimani*[1,3*†]

[1] *Dipartimento di Scienze biomediche e Sanità pubblica, Università Politecnica delle Marche, Ancona, Italy,* [2] *Dipartimento di Comunicazione ed Economia, Università degli Studi di Modena e Reggio Emilia, Reggio Emilia, Italy,* [3] *Munich Center for Mathematical Philosophy, Ludwig-Maximilians-Universtät München, München, Germany*

OPEN ACCESS

Edited by:
Cedric Bousquet, Centre Hospitalier Universitaire (CHU) de Saint-Étienne, France

Reviewed by:
David Madigan, Columbia University, United States
Pietro Panei, National Institute of Health (ISS), Italy

***Correspondence:**
Barbara Osimani
b.osimani@univpm.it

†ORCID
Francesco De Pretis
orcid.org/0000-0001-8395-7833
Jürgen Landes
orcid.org/0000-0003-3105-6624
Barbara Osimani
orcid.org/0000-0001-5212-9525

Specialty section:
This article was submitted to Pharmaceutical Medicine and Outcomes Research, a section of the journal Frontiers in Pharmacology

Received: *14 February 2019*
Accepted: *15 October 2019*
Published: *17 December 2019*

Citation:
De Pretis F, Landes J and Osimani B (2019) E-Synthesis: A Bayesian Framework for Causal Assessment in Pharmacosurveillance. Front. Pharmacol. 10:1317. doi: 10.3389/fphar.2019.01317

Background: Evidence suggesting adverse drug reactions often emerges unsystematically and unpredictably in form of anecdotal reports, case series and survey data. Safety trials and observational studies also provide crucial information regarding the (un-)safety of drugs. Hence, integrating multiple types of pharmacovigilance evidence is key to minimising the risks of harm.

Methods: In previous work, we began the development of a Bayesian framework for aggregating multiple types of evidence to assess the probability of a putative causal link between drugs and side effects. This framework arose out of a philosophical analysis of the Bradford Hill Guidelines. In this article, we expand the Bayesian framework and add "evidential modulators," which bear on the assessment of the reliability of incoming study results. The overall framework for evidence synthesis, "E-Synthesis", is then applied to a case study.

Results: Theoretically and computationally, E-Synthesis exploits coherence of partly or fully independent evidence converging towards the hypothesis of interest (or of conflicting evidence with respect to it), in order to update its posterior probability. With respect to other frameworks for evidence synthesis, our Bayesian model has the unique feature of grounding its inferential machinery on a consolidated theory of hypothesis confirmation (Bayesian epistemology), and in allowing any data from heterogeneous sources (cell-data, clinical trials, epidemiological studies), and methods (e.g., frequentist hypothesis testing, Bayesian adaptive trials, etc.) to be quantitatively integrated into the same inferential framework.

Conclusions: E-Synthesis is highly flexible concerning the allowed input, while at the same time relying on a consistent computational system, that is philosophically and statistically grounded. Furthermore, by introducing evidential modulators, and thereby breaking up the different dimensions of evidence (strength, relevance, reliability), E-Synthesis allows them to be explicitly tracked in updating causal hypotheses.

Keywords: adverse drug reaction, drug safety, causal assessment, Bradford Hill Guidelines, statistical evidence, evidence synthesis, evidence quality, pharmacovigilance

BACKGROUND

The United States Department of Health and Human Services reports that although medications help millions of people live longer and healthier lives, they are also the cause of approximately 280,000 hospital admissions each year and an estimated one-third of all adverse events in hospitals (US Department of Health and Human Services, Office of Disease Prevention and Health Promotion, 2014). The problem of adverse drug reactions is obviously not confined to the USA, but is a global issue (Edwards and Aronson, 2000; European Commission, 2008; Wu et al., 2010; Stausberg and Hasford, 2011). Evidence facilitating the prediction of adverse drug reactions often emerges unsystematically and unpredictably in the form of anecdotal reports, case series, and survey data, as well as more traditional sources, e.g., clinical trials (Price et al., 2014; Onakpoya et al., 2016). Recently, legislators have called for the integration of information coming from different sources when evaluating safety signals (European Parliament and the European Council: Directive 2010/84/EU; Regulation (EU) No 1235/2010; see also the 21st Century Cures Act, recently entered into force in the US). A similar call has also been issued by researchers (Cooper et al., 2005, p.249) and (Herxheimer, 2012). However, standard practices of evidence assessment are still mainly based on statistical standards that encounter significant difficulties with the integration of data emerging from observational and experimental studies at times on different species as well as from lab experiments and computer simulations. Clearly, there is increasing awareness of the need for tools that support the assessment of putative causal links between drugs and adverse reactions grounded on such heterogeneous evidence.

Indeed, the body of methodological work on post-marketing risk management *via* the aggregation of evidence is rapidly growing. The recent focus has been on various aspects of causal assessment based on heterogeneous evidence. Some examples include work on aggregating human and animal data (European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC), 2009), aggregation of spontaneous reports (Caster et al., 2017; Watson et al., 2018), Bayesian aggregation of safety trial data (Price et al., 2014) and data sets (Landes and Williamson, 2016), bringing together toxicology and epidemiology (Adami et al., 2011), retrieving but not assessing evidence Knowledge Base workgroup of the Observational Health Data Sciences and Informatics, 2017; Koutkias et al., 2017), assessing the evidential force of data in terms of reproducibility and replicability of the research (LeBel et al., 2018), grading certainty of evidence of effects in studies (Alonso-Coello et al., 2016), grading observational studies based on study design (Sanderson et al., 2007; Sterne et al., 2016; Wells et al., 2018), thematic synthesis of qualitative research, decision making (Thomas and Harden, 2008; Landes, 2018), providing probability bounds for an adverse event being drug induced in an individual (Murtas et al., 2017) in Pearl's formal framework for causality (Pearl, 2000) and work on aggregating evidence generated by computational tools (Koutkias and Jaulent, 2015).

Much work has been devoted to the development of evidence synthesis methods testified by a growing number of (systematic) reviews and comparisons of evidence synthesis methods (Lucas et al., 2007; Greenhalgh et al., 2011; Kastner et al., 2012; Warren et al., 2012; van den Berg et al., 2013; Tricco et al., 2016a; Tricco et al., 2016b; Kastner et al., 2016; Shinkins et al., 2017). A number of studies argue that while there are many approaches and standards, it is not at all clear which is best (Greenhalgh et al., 2011; Warren et al., 2012; van den Berg et al., 2013; Kastner et al., 2016; Tricco et al., 2016a; Tricco et al., 2016b).

Traditional approaches supporting drug-licensing decisions are reviewed in (Puhan et al., 2012) and the changing roles of drug-licensing agencies in an evolving environment are described in (Ehmann et al., 2013). Closest to our approach are those that employ Bayesian statistics (Sutton and Abrams, 2001; Sutton et al., 2005).

However, the number of approaches that attempt to tackle the issues of aggregating different types of evidence to facilitate causal assessment of adverse drug reactions (assessing whether a drug causes an adverse reaction) straight on is rather small. One such approach is an epistemological framework based on Bradford Hill's well-known guidelines (Hill, 1965), which continue to be an active area of research, e.g., see (Swaen and van Amelsvoort, 2009; Geneletti et al., 2011; Fedak et al., 2015).

Our work is rooted in the tradition that draws on statistical information and probabilsitc (in)dependence for the purpose of causal assessment. Challenges to Bayesian causal assessment have been raised by Dawid et al. (2016) among others.

This paper is a first step towards translating the philosophical approach to causal assessment of suspected adverse drug reactions of (Landes et al., 2018) towards an applicable framework. The rest of the paper is organised as follows. Next, we introduce and expand the approach of (Landes et al., 2018) and build a Bayesian network model for it. Then we apply the framework and model to a case study and conclude.

Here, we are mainly concerned with further developing the framework and how to – in principle – operationalise our approach. Delineated functional forms and some (conditional) probabilities serve only illustrative purposes. The focus is on how to determine them in principle and highlight roles and interactions of relevant concepts. Hence, significant further work is required before the framework is a ready-to-use tool.

METHODOLOGY

E-Synthesis is a theoretical framework for causal assessment based on (Landes et al., 2018), we briefly present here its main components and integrate further dimensions of evidence.

Aims and Scope

The framework in (Landes et al., 2018) aims to support decision making in drug regulatory agencies by providing a probability that a drug causes an adverse reaction.[1]

[1]Note that the framework does not aim to provide utilities of harms, nor probabilities of expected benefits, nor utilities of benefits. When utilities of harms and benefits, as well as estimation of benefits are provided by further means, a drug regulatory agency can perform an expected utility calculation to determine whether the expected advantages of a drug exceed the expected dis-advantages. The agency will withdraw the drug (or not approve it), if the expected disadvantages outweigh the expected benefits, see (Landes et al., 2018, *Bayesian Network Model*).

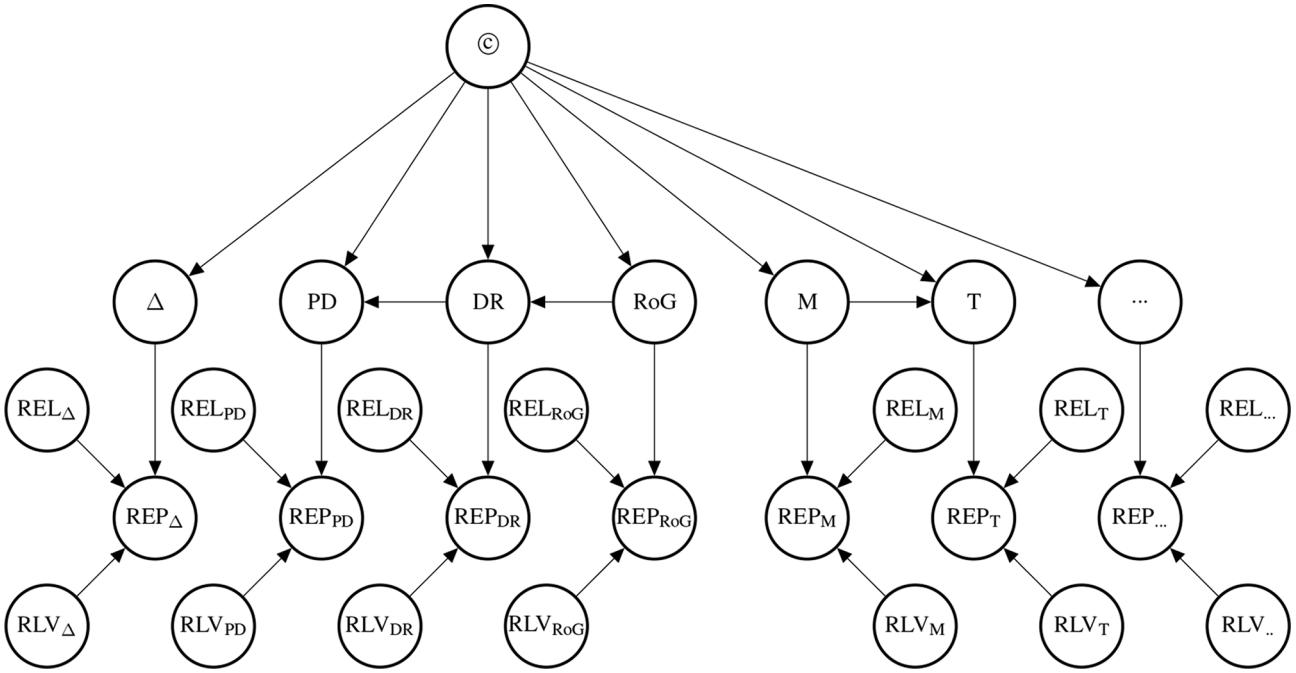

FIGURE 1 | Graph of the Bayesian network with one report for every causal indicator variable taken from Landes et al. (2018). The dots indicate that there might be further indicators of causality not considered here. As explained in text, we here take it that *M* (mechanistic knowledge) entails *T* (temporal precedence) and hence introduce an arrow from *M* to *T* which is not in Landes et al. (2018). In the original paper, we considered two modulators *REL* (for "reliability" of study authors) and *RLV* [for "relevance" (external validity)] act as evidential modulators of reports (the *REP*-nodes). In this paper, we focus exclusively on the *REL* modulator which we split into a number of concepts, see *Evidential Modulators: Study Design*.

The hypothesis of interest is that "Drug $\mathcal{D}$ causes harm E in population U." To facilitate the inference from all the available evidence, indicators of causality are used. These indicators are based on Hill's nine viewpoints for causal assessment (Hill, 1965).

Bayesian Network Model

The probability of the causal hypothesis © is modelled *via* a Bayesian network of a finite number of propositional variables, see (Landes et al., 2018, *Discussion*) and see (Neapolitan, 2003) for a standard introduction to Bayesian networks. There is a binary propositional variable for the hypothesis of interest and a binary propositional variable for all six indicators of causality. For every item of evidence, every source of evidence and every study population, we create a report and evidential modulators, see **Figure 1**. The conditional probabilistic (in-)dependencies can be gleaned from the graph in terms of (Pearl, 2000)'s *d*-separation criterion. Although, the Bayesian network is used for causal assessment, the arrows in the network are *not* causal arrows in (Pearl, 2000)'s sense – they here represent epistemic probabilistic (in-)dependencies only. **Figure 2** is an example graph with only one report. The report is the only child of many parents.

The causal hypothesis represented by © is a root node.[2] The causal indicator are parents of report nodes which mediate causal inference from the concrete data (the reports) towards the causal hypothesis. The parents at the two bottom levels are modulators of the evidential strength of the data. These incorporate considerations about the reliability of the evidence into the assessment of the hypothesis. In particular, they take into account the possibility of random error [as a function of sample size (*SS*) and study duration (*D*)], and systematic error; attenuated by adjustment or stratification (*A*), randomisation (*R*), blinding (*B*), placebo (*Pl*) and sponsorship bias (*SB*).

This framework also allows for the incorporation of evidential modulators related to external validity [called "relevance" in (Landes et al., 2018)], however we will not treat them here for ease of exposition.[3]

A probability function consistent with the conditional independencies of the Bayesian network is selected which expresses our uncertainties in the tradition of Bayesian epistemology (Bovens and Hartmann, 2003; Howson and Urbach, 2006; Talbott, 2011). Unlike in "pure" Bayesian statistics, where one conditionalises on statistical models and hence obtains conditional probabilities mandated by the particular model (parameter), in Bayesian epistemology one may conditionalise on any proposition (or event), since probabilities are interpreted more widely as one's uncertainties about general propositions; the Bayesian statistician Lindley is sympathetic to this approach (Lindley, 2000). In case one does conditionalise on a particular

[2]The abbreviations used in this article are listed in **Table 1**.

[3]They are extensively addressed in (Poellinger, 2019) and will be the focus of a rejoinder to this paper.

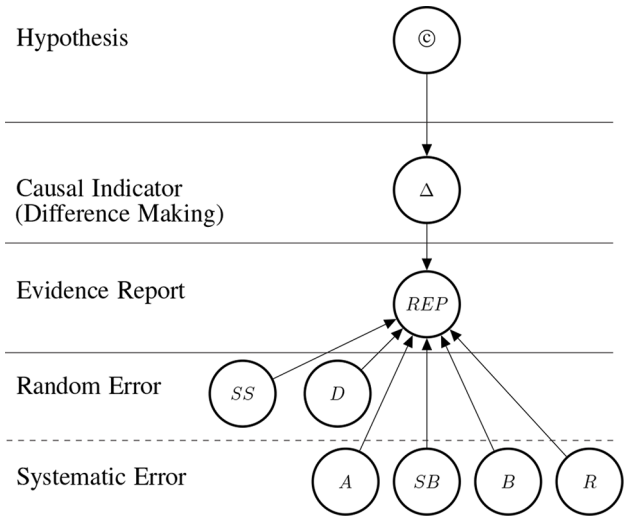

FIGURE 2 | Graph structure of the Bayesian network for one RCT (randomized controlled trial) which informs us about difference making (Δ) which in turn informs us about the causal hypothesis. The information provided by the reported study is modulated by how well the particular RCT guards against random and systematic error. Duration affects systematic and random error as explained in *Control for Random Error*.

model, conditional probabilities are (virtually always) set to the probabilities of the statistical model.

While a certain degree of subjectivity is undeniable, there is a good argument to be made that some subjectivity is unavoidable in any approach to statistical/uncertain inference (Gelman and Hennig, 2017) and that Bayesian epistemology is in fact objective; or objectivity conducing; to some degree (Sprenger, 2019).[4]

Theoretical Entities

Concepts of interest fall into two classes: i) a class of causal concepts comprising the hypothesis of interest and the indicators of causation and ii) a class of evidential concepts comprising evidential modulators and reports (data).

The Causal Hypothesis (©)

We are interested in determining the probability of the causal hypothesis that a drug $\mathcal{D}$ causes a particular adverse effect E in a population U – given the available evidence. Although the hypothesis space could be in principle subdivided into three hypotheses: 1) $\mathcal{D}$ causes E, 2) $\mathcal{D}$ hinders E, and 3) $\mathcal{D}$ does not cause E, we divide it here for simplicity's sake into two alternative hypotheses: 1) $\mathcal{D}$ causes E and 2) $\mathcal{D}$ does not cause E, which consists of the disjunct of 2 and 3 above. To shorten notation, we use the symbol © in order to denote causation, such as in $\mathcal{D}©E$, or simply ©.[5]

TABLE 1 | Abbreviations.

Symbol	Intended Interpretation
A	Adjustment for Confounders
avg	Average
B	Blinding
CI	Confidence Interval
$\mathcal{D}$	Drug
D	Duration
DR	Dose-response Relationship
ES	Effect Size
M	Mechanistic Knowledge
M_i	Mechanistic Hypothesis
NAQPI	N-acetyl-p-benzoquinone imine
OR	Odds Ratio
PD	Probabilistic Dependence
PI	Placebo
R	Randomisation
RCT	Randomized Controlled Trial
Rep	Report Variable
RoG	Rate of Growth
SB	Sponsorship Bias
SS	Sample Size
ST	Signal-Tracking
T	Temporal Precedence
TRPA1	Transient Receptor Potential Ankyrin-1
©	Hypothesis of Causation
Δ	Difference Making
μ_i	Mechanistic Report Variable
$\mathbb{R}$	Set of Real Numbers
Σ	Set of Statistical Indicators: *PD, DR* and *RoG*

Indicators of Causation

Causal inference is mediated in the framework by "indicators of causation" in line with the Bradford Hill Guidelines for causation. As Hill puts it (Hill, 1965):

> "None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a sine qua non. What they can do, with greater or less strength, is to help us make up our minds in the fundamental question – is there any other way of explaining the set of facts before us, is there any other equally, or more, likely than cause and effect?"[6]

In epistemic terms, causal indicators can be considered as observable and testable consequences of causal hypotheses, albeit non-deterministic consequences (with one exception); that is, they are more likely to be observed in the presence of a causal relationship and less likely in its absence, $P(Ind\,|\,©) > P\left(Ind\right) > P(Ind\,|\,\overline{©})$ but they are not entailed by it.

The first indicator "difference-making," Δ, is a perfect one, in that it entails causation. However, note that in our framework Δ is not entailed by causation. All other indicators are related

[4]What probabilities at a population level may mean to an individual has recently been explored in (Dawid, 2017).

[5]We do not commit here to any specific view or definition of causation [e.g., dispositional, probabilistic, counterfactual, manipulationist, etc. see also (Landes et al., 2018)]. That's a question on the ontology of causation that we leave open for the moment. Our causal hypothesis allows for the term "causes" to cover any of the current definitions of causality to the extent that the evidence used for causal inference may be made relevant to them.

[6]Bradford Hill both refers to explanatory power and likelihood as reliable grounds to justify causal judgements, and presents the respective criteria as opposed to tests of significance: "No formal tests of significance can answer those questions. Such tests can, and should, remind us of the effects that the play of chance can create, and they will instruct us on the likely magnitude of those effects. Beyond that, they contribute nothing to the proof of our hypothesis." (Hill, 1965).

only probabilistically to the hypothesis of causation, as we now explain.

Difference-Making (Δ)

If $\mathcal{D}$ and E stand in a difference-making relationship, then changes in $\mathcal{D}$ make a difference to E (while the reverse might not hold). In contrast with mere statistical measures of association, the difference-making relationship is an asymmetric one. Probabilistic dependence can go in both ways (e.g., if Y is probabilistically dependent on X, then also X is probabilistically dependent on Y); the same does not hold for difference making, which provides information about its direction. This explains why experimental evidence is considered particularly informative with respect to causation; the reason is exactly that in experiments, putative causes are intervened upon, in service of establishing whether they make a difference to the effect.[7]

Consistent with our choice of modelling "positive" causation only (that is instances of causation where X fosters rather than inhibiting Y), we shall understand this difference-making indicator as being true, if and only if the difference made is a positive one. *Mutatis mutandis*, this convention applies to the following three indicators as well.

Probabilistic Dependence (PD)

PD encodes whether $\mathcal{D}$ and E are probabilistically dependent or not – such dependence naturally increases our belief in some underlying causal connection (as an indicator of causation; see, e.g., Reichenbach, 1956). Probabilistic dependence is an imperfect indicator of causation because neither entails the other. There are cases in which probabilistic dependence is created by confounding factors, as well as cases where two opposite effects of a single cause cancel each other out and produce a zero net effect.[8]

Dose-response Relationship (DR)

Dose-response relationships are taken as strong indicators of causation. *DR* is a stronger indicator than probabilistic dependency alone, because it requires the presence of a clear pattern of ≥3 data-points relating input and output. Indeed *DR* implies *PD*. Dose-response relationships can be inferred both at the population and at the individual level, and both in observational and experimental studies. Dose-response curves correspondingly have different scopes (e.g., the time-trend coincidence of paracetamol purchase and asthma increase in a given population [(Newson et al., 2000) vs. clinical measurements of concentration effects of analgesics]. *DR* abstracts away from these specifications and means that for dosages in the therapeutic range, the adverse effect E shows (approximate) monotonic growth for a significant portion of the range (see below, **Figure 3**, for an illustration of important types of dose-response curves).

Rate of Growth (RoG)

This indicator refers to the presence of a steep slope in the dose-response relationship. Hence, *RoG* implies a dose-response relationship (*DR* without *RoG* means either that the rate of growth is low, or highly non-linear). The indicators of causality *RoG*, *DR*, *PD* are independent of the causal structure, in the sense that they could be equally observed either in cases where $\mathcal{D}$ causes E, or in cases where E causes $\mathcal{D}$, or when $\mathcal{D}$ and E have a common cause. All that matters is whether there is a (certain) systematic relationship between $\mathcal{D}$ and E. *RoG*, *DR*, *PD* are semantically and epistemically related and we refer to them as statistical "black-box" indicators, denoted by Σ.

Mechanistic Knowledge (M)

M represents the proposition: "there is a mechanism for $\mathcal{D}$ to E": by which we mean a "linkage between a direct molecular initiating event [..] and an adverse outcome at a biological level of organization relevant to risk assessment" (Ankley et al., 2010, Page 731). In the biological realm, a causal relationship obviously entails the presence of a biological mechanism connecting the

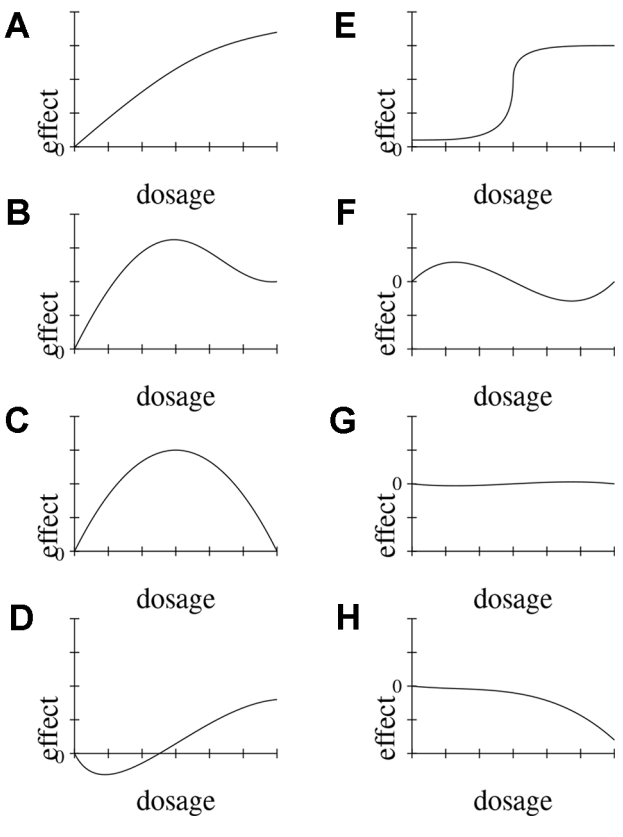

FIGURE 3 | Possible functional forms of the relationship between dosage and effect. We delineate eight **(A–H)** exemplary functional forms for this relationship.

[7]In philosophical terms, difference-making is understood as *ideal controlled variance* along the concept of *intervention* in manipulationist theories of causation (see Landes et al., 2018 for a detailed treatment, see also Woodward, 2003 and Pearl, 2000)): X is called a cause of Y if Y's value can be varied by varying X (possibly upon controlling for additional variables in the given situation).

[8]A well-known example of this type of cancellation is Hesslow's birth control pills case (see, e.g., Cartwright, 2001): The contraceptive (directly) causes thrombosis but simultaneously (indirectly) prevents thrombosis by preventing pregnancy which is a cause of thrombosis. (Cartwright, 2001) discusses this case as one of the pitfalls of reducing causal analysis to probabilistic methodology alone. Of course, if cancellation is suspected, one might disable certain preventative causal routes to check whether the causal relationship actually shows once disabling conditions are held fixed. Cartwright however discusses cases where this strategy might not even be viable, owing to the complexity of the causal web.

cause to the effect. Therefore, © ⇒ *M*. However, a mechanism may not be causally responsible for bringing out the effect due to possible inhibitors, back-up mechanisms, feedback loops, etc. *M* ⇒ © does hence not necessarily hold.

Time Course (T)

T encodes whether $\mathcal{D}$ and *E* stand in the right temporal relationship (time course), which can refer to temporal order, distance, or duration. If $\mathcal{D}$ causes *E*, *T* must hold (as a necessary condition): © ⇒ *T* . *T* remains an imperfect indicator, nevertheless, because temporal precedence is also compatible with ¬($\mathcal{D}$ causing *E)* when $\mathcal{D}$ and *E* are connected by a common cause or through reversed causation. Hence *T* ⇒ © does not necessarily hold.

Relationships Between Causal Indicators and the Causal Hypothesis

As mentioned above, since they are observable consequences of ©, the causal indicators (Δ, *PD*, *DR*, *RoG*, *M*, *T*) provide support for the causal hypothesis © once concrete studies provide concrete evidence for these indicators. **Figure 4** illustrates the conceptual dependencies among these indicators (in Bayes net style). As mentioned above, not all indicators have the same strength: Δ is understood as a perfect indicator (Δ ⇒ ©) of causality. However, because of the possibility of holistic causation the inverse does not hold, that is, it is not the case that ©⇒Δ. With holistic causation we refer to cases in which causal links are embedded in a causal structure, which does not allow surgical interventions on the individual causal links (see Cartwright, 2007; Mumford and Anjum, 2011).

Landes et al. (2018) explain that although we tend to identify causation with a systematic and, possibly, asymmetric relationship between two entities or variables, yet, we prefer, in the context of causal inference, to remain neutral towards the various definitions of causation provided in terms of necessary and sufficient conditions in the philosophical literature. We choose to allow for "weaker" markers of causation, such as imperfect indicators[9] rather than requiring the satisfaction of necessary and/or sufficient conditions of causation. However, since the indicators of causation are weaker versions of the requirements for causation formulated in the philosophical literature, this framework may be considered to generalise these requirements and the distinct ways to formulate them. A possible user is allowed to adopt (exclusively) any one of them, or to assume a pluralistic stance thereby benefiting from various inferential paths.

One reason for not having © ⇒ Δ is precisely to allow our framework to incorporate holistic conceptualisations of causation in contrast to the modular conceptualization of causes typical of the causal graph methodology developed by Pearl (2000) and Glymour and colleagues (Spirtes et al., 2000) see also Woodward (2003).[10]

[9]That is, indicators that merely make the probability of a causal relationship being present more probable than not.

[10]These approaches are under attack for failing to recognize that causal structures may be holistic (that is, synergistic and non-additive vs. modular and additive), and therefore may be not adequately captured by a difference making account. Strictly speaking this sort of criticism does not deny that Δ ⊃ ©, but only denies the reverse: © ⊃ Δ. However, in the causal graph literature the defining features for causality jointly entail that Δ ⇔ ©. We respect this debate by not collapsing Δ and © into a single node.

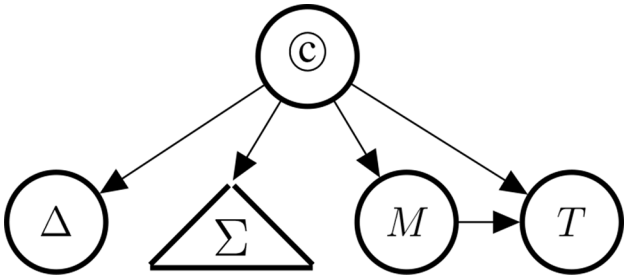

FIGURE 4 | Indicators of causality, clustered by type: Δ (difference-making), Σ (the statistical blackbox) consisting of *PD* (probabilistic dependence), *DR* (dose response), and *RoG* (rate of growth), *M* (the existence of a mechanism), and *T* (temporal precedence).

The presence of a high rate of growth, *RoG*, in the dose-response relationship supports causation more strongly than the dose-response *DR* would by itself without being steep. *PD* is the weakest indicator in the Σ set. Note, however, that the statistical concepts are unrelated to difference-making information, if we have knowledge about the causal link (we will make this very fact explicit in Bayes net terms below). This reflects our intention to demarcate the conceptual divide between purely observational (symmetric) and genuinely interventional (asymmetric) indicators. Moreover, *M* entails *T* in that if there exists a mechanism linking the drug and the side effect, then it must be the case that drug administration and side effect stand in the right temporal order.

Evolution of Our Approach

The model in (Landes et al., 2018) was developed to formalise causal inference in pharmacology on a fundamental level. It lacks the complexity necessary to capture certain important aspects of practical applications. For example, all variables are binary. Furthermore, studies are either deemed unreliable and do not provide any information whatsoever or they are deemed fully reliable and thus prove or disprove causal indicators. Conditional probabilities of causal indicators were left unspecified. Mechanistic evidence was not given particular attention.

In this paper, we allow for continuous variables taking values in the entire unit interval [0,1] ⊂ ℝ, discuss and model in detail the inferential roles of evidential modulators and thereby improve on the model of reliability (*Evidential Modulators: Study Design* and **Supplementary Material**), give a method for determining conditional probabilities of causal indicators (**Supplementary Material**) and show how mechanistic reasoning may be formalised (see Sections devoted to Mechanistic Evidence in the theoretical part and in the case study, and Section 4 of the **Supplementary Material**). In *Application of the Model: Does Paracetamol Cause Asthma?*, we show how the current model can be applied to the debated causal connection between paracetamol and asthma.

The ultimate goal is to evolve our philosophical perspectives on causal inference into a ready-to-use instrument for causal assessment supporting actual decision making procedures. This paper constitutes a step in this direction.

Evidential Modulators: Study Design

In analogy to (Bogen and Woodward, 1988), we split the inferential path into two stages, one leading from data to abstract phenomena (here, causal indicators), and one from such phenomena to theoretical entities (here, causation). This allows us to distinguish theoretical issues related to causation and their consequences for the purpose of causal inference, from methodological concerns associated with the interpretation of data. At this second stage, we model the signal-tracking ability of the reports as a function of the instrument (the study) with which the evidence was gathered.

The signal-tracking depends on how much the study design is supposed to have controlled for systematic and random error, that is minimisation of the chances that a causal effect is wrongly attributed to the treatment under investigation, when instead the effect is due to other factors or to chance (false positive), or vice versa (false negative). Indeed, a plausible interpretation of the criterion underpinning the evidence hierarchies is the maximisation of internal validity, see also (La Caze, 2009).

However, in our view, study design also determines the kind of information that the evidence is able to provide, hence we evaluate the evidence also on the level of the kind of information it delivers: that is, the causal indicator it is able to "speak to."

Following point 1, we associate distinct types of study design as potential carriers of causal indicators as follows:

1. Randomised Controlled Trials (RCTs) provide information about difference making, time course, possibly also dose-response relationship and rate of growth.
2. Cohort studies provide evidence of time course and statistical association (Σ).
3. Case-control studies provide information about Σ only.
4. Individual case studies cannot provide information about statistical association, but they provide very detailed information about time course and, possibly, difference-making, whenever this can be established with confidence [see for instance the Karch-Lasagna or Naranjo algorithm (Karch and Lasagna, 1977; Naranjo et al., 1981; Varallo et al., 2017)]. However, they provide very local information, about an individual subject, and therefore do not license inferences about the general population.
5. Case series can possibly help delineate a reference class, where the putative causal link holds.
6. Basic science studies (*in vitro*, or in silico), and *in vivo* studies, are generally the main source for evidence on the mechanisms underpinning the putative causal link.

The distinction of different dimensions of evidence, beyond different lines of evidence, and different inferential levels (main hypothesis, indicators, data, modulators) is the innovative point of our approach with respect to the standard view, in which these aspects are conflated, or, at least, remain implicit, in evaluating and using evidence in order to make decisions. The reason for adopting such an approach is twofold:

1. To avoid conflation of distinct ways in which the available evidence bears on the hypothesis of interest. Among others, this characterization, makes more explicit what distinguishes one method from another in terms of relevant causal information, rather than of the degree to which it avoids systematic error;
2. In our framework, evidence supports the causal indicators which in turn support the causal hypothesis of interest, each to a different degree. Hence, downgrading the evidential value of studies that feed into the weaker indicators just because of the kind of information they cannot provide, would amount to double-downgrading such evidence. For instance, evidence coming from observational studies is uninformative with respect to the Δ indicator, but may be highly informative with respect to statistical association (Σ).

Therefore, the kind of study from which the evidence derives is directly specified by the kind of indicators to which it speaks, which be found on the right side of the conditional probability (see Section 2 in the **Supplementary Material**).[11]

Additionally, studies are weighted by their degree to which they control for systematic and random error. Control for random error is operationalised in terms of Sample Size (*SS*) and Study Duration (*D*). Control for systematic error is operationalised differently for experimental vs. observational studies. We assume that for pure observational studies, signal-tracking is limited to getting the *statistical* indicators right. We consider adjustment/stratification as relevant procedures in this respect.

Instead, for experimental studies, signal-tracking relates to getting the *causal link* right. Therefore, control for systematic error also includes attributes connected to excluding alternative causal explanation for the observed effect, such as blinding, and randomisation. In both cases we add an indication of whether the study could have been intentionally biased (because of financial interests). In the following, we discuss these evidential modulators in more detail.

In the future, we hope to analyse and incorporate further modulators such as dropouts, missing data, protocol violations, whether analysis was by intent to treat and the presence or absence of further biases into our approach. For example, regarding harm assessment, which is the focus of the present study and the main goal for developing *E-Synthesis*, sponsorship bias shifts probabilities towards reports of greater safety. (Sub-conscious) biases may instead push researchers in both directions, with a higher prevalence towards reporting more publishable results: this means statistically significant evidence and/or counter-intuitive and surprising results.

Our Bayesian model is sufficiently powerful to capture uncertainties arising from inherent difficulties in assessing the degree to which studies are controlled for systematic and random error.

Control for Random Error

Sample Size (SS)

A large sample size helps to reduce confidence in the hypothesis that an observed effect (or lack thereof) is due to chance/noise/

[11]For instance, $P(ES \mid A, S, D, SB, PD, T)$ is the conditional probability of observing blue whether there is an effect size *ES* or not in a cohort study; where the instantiated indicators on the right side of the conditional probability are *PD*, and theindicator for time course (*T*), since cohort studies provide information about both indicators (that is evidence for or against each of them).

random error. The larger the sample size, the less defeasible the inference one may draw from reported results (modulo systematic error).

Study Duration (D)

Most drugs produce their beneficial effects within a time horizon that is well-understood at the time of drug prescription. Instead, some adverse drug reactions, such as stroke, heart attack, and cancer, may be noted only a long time after the end of the treatment. A priori, it is not clear after how long the adverse effects will materialise. Infamous examples are the DES tragedy of causing vaginal adenocarcinoma in pubertal and adult children of treated pregnant mothers (Preston, 1988) and antipsychotic drugs causing tardive dyskinesia after years of treatment (Beasley et al., 1999).

In principle, the longer the follow up, the more likely adverse drug events will be detected. Studies with a short follow up period may thus fail to detect medium to long term effects of drugs, hence they tend to produce false negatives. A study with a short follow-up period, which does not detect an adverse effect, can only count as very weak evidence against the causal hypothesis, since the adverse reaction may occur only after the end of the follow-up period (see Vandenbroucke and Psaty, 2008). However, if the drug does not cause an adverse effect, then the study duration does obviously not influence the probability of finding it in the studied population.[12]

So, study duration affects *random* error but short studies lead to a *systematic* under-reporting of harms. This explains the position of the duration node in **Figure 2**.

Control for Systematic Error

While large sample sizes and long-term studies allow one to reduce one's belief in a chance result, one has thereby *not excluded* other factors that may have caused the results. For example, consider a large study that is biased in an important respect, then – when evidence is taken at face value – one may become even surer that one has nailed down the effect size of the phenomenon of interest, erroneously so (this bias tends to become "intransigent" the larger the sample size becomes; see: Holman and Bruner, 2015). In 1998, the point was made thus: "*There is a danger that meta-analyses of observational data produce very precise but equally spurious results*" (Egger et al., 1998, p. 140). This point has recently been explored in computer simulations for aggregating evidence *via* frequentist statistics (Romero, 2016).[13]

Blinding, Randomisation and Placebo (B,R,Pl)

The main instruments to isolate the putative causal link $\mathcal{D} \copyright E$ from all other possible causal effects on E beyond chance. This happens because, through randomisation, one has a probabilistic guarantee (modulo random error), that the treatment and the control group are comparable with respect to all these possible additional causal influences, and therefore that the observed effect is due to the treatment and only to it, see (Fuller, 2019) for a philosophical discussion. Double blinding ensures that randomisation is not biased by the researcher in order to obtain a wishful result, or by the study subjects, through so-called placebo effects.

When the experiment has no placebo arm, or none of the drugs in the control arms are sufficiently understood, then the study cannot deliver any information about Δ (and not even about any of the Σ indicators), since evidence for these indicators draws on the observed effect difference *with* and *without* the presence of the putative cause.

In fact, if the effects of distinct *putative* causes are compared against each other, without any knowledge as to their causative status, and no absolute benchmark (i.e. *absence* of all putative cause), then such relative comparison against each other only provides information about relative difference making, that is, one is not able to establish whether e.g. 1) drug $\mathcal{A}$ produces an improvement of symptoms, 2) or it is drug $\mathcal{B}$ that worsens the situation by the same amount, 3) or else, both drug $\mathcal{A}$ and $\mathcal{B}$ have opposite effects with respect to such symptoms (the former improves them, while the latter worsens them).

There are cases that can be disambiguated though, for instance when at least one of the arms but not all of them show a dose-response relationship. In this case, the very fact that some drugs do not exhibit such dose-response relationship, and some do, is taken as a sign that the latter do contribute to E in some way, while the former can be taken as benchmark(s).

Even when no such disambiguation is feasible, there is still a possibility to glean some information about Δ in experimental studies, by comparing the study outcomes to the base rate incidence of the same outcome measure in the sampled population. Like many other steps in causal inference this step is fraught with risks. The more tenuous the connection between study outcome and the base rate, the more risky the step. In our framework, this risk is captured by employing different values of the evidential modulator representing the quality of the implementation of placebo control.

Adjustment and stratification (A)

Both in experimental and observational studies data may be adjusted for covariates both in the design and in the analysis phase. This may be done in various ways: factorial design, stratification, standardisation, multivariate regression analysis, and, more recently with the aid of Propensity Score methods (Montgomery et al., 2003; Kurth et al., 2005; Schneeweiss et al., 2009; Kahlert et al., 2017). This is an important attribute in the methodology of causal inference, which is however fraught with several diagnostic pitfalls, especially due to the requirement of "causal sufficiency" (any causal inference is invalidated, if the set of covariates on which it is based misses latent variables). Adjusting *for the right covariates, in a sufficient causal set* leads us to detecting non-spurious statistical associations, whereas conditioning on the wrong variables leads us astray and increases

[12]A study with a long follow up may have measured safety (end-) points only at the end of the study or at multiple times. That's a relevant consideration concerning study duration not yet incorporated here.

[13]However, the probability of imbalance between the two groups decreases with increasing N. Hence, very large samples, especially with random sampling, may guard not only against random error, but also against some forms of confounding.

the chance of false positives and negatives, see, e.g., (McCarron et al., 2010).[14]

Sponsorship Bias (SB)

Evidence hierarchies are one means to order study designs in terms of the potential for suffering from systematic error, either caused by confounding or by intentional distortion of the evidence. While higher level evidence – RCTs, meta-analyses, systematic reviews of meta-analyses - is in principle less manipulable (because of blinding, randomisation, and increased accuracy through data pooling), still, well-known incentives to the distortion of evidence may arise through vested interests, and compromise the reliability of the evidence at different stages of evidence collection, interpretation and evaluation quite independently of the methodology adopted (Rising et al., 2008; Wood et al., 2008; Song et al., 2009; Krauth et al., 2013; Ioannidis, 2016). Other things being equal, a sponsored study is more likely to produce results which align with the sponsor's interest.

One persistent bias in medical research is the sponsorship bias due to the interests of the organisations funding medical research, see, e.g., (Lundh and Bero, 2017; Lundh et al., 2017). One dramatic instance of sponsorship influencing the safety evaluation of a drug is the Vioxx disaster (Jüni et al., 2004; Horton, 2004).[15] If a drug causes an adverse reaction, a sponsorship bias tends to hide it, and therefore makes it more likely that the study delivers no reports about adverse drug events (or reports with smaller effect sizes than the drug really induces). Furthermore, by tending to distort results in a predefined direction, bias "interacts" with random error, in the sense that systematically biased procedures, when replicated, lead to increased "artificial" accuracy: it may well be that the rate of false negatives is higher for non-sponsored studies, hence apparently paradoxically, sponsorship *bias* produces "more accurate" data when no side-effects are present in reality.[16]

Regulatory constraints on medical methodology have evolved with such sources of bias in mind, see (Teira, 2013; Teira and Reiss, 2013). However, as some have recently noted, those who intend to manipulate data find ways circumventing such regulatory constraints and trigger a race of arms characterised by epistemic asymmetry (Holman, 2015; Holman and Geislar, 2019).[17]

Reports

This section lists three possible evidence types which may be observed with respect to causal hypotheses in medicine. A certain (statistical) measure as to the effect size, evidence regarding the possible mechanisms underpinning the "phenotypic" effect, and evidence of time course (which can only come jointly with one of the other two).[18]

Effect Size (ES)

The medical community has developed various popular measures of the strength of observed effects: relationships between the odds ratio, hazard ratio and the relative risk are discussed in (Stare and Maucort-Boulch, 2016; Sprenger and Stegenga, 2017).[19]

These measures all refer to the average observed effect difference in the study groups. However, other measures of causal strength refer to the systematic pattern that relates treatment and effect (dose-response relationship) and to the rate at which increase in dosage increases the observed effect (rate of growth).

Mechanistic Evidence (ME)

Evidence speaking for or against a mechanistic hypothesis stems from basic science or animal studies, and previously established pharmacological/biochemical knowledge. It is rarely the case that a study confirms or establishes a complete mechanism of action ["a complete and detailed understanding of each and every step in the sequence of events that leads to a toxic outcome" European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC), 2007, Page 13)] by which a drug causes an adverse reaction. Instead, mechanistic knowledge is most often acquired piecemeal: incoming evidential reports are put together to complete a mechanistic puzzle, and they acquire their meaning only within the broader picture.

Time Course

Evidence of time precedence can come from experimental studies (e.g., RCTs), from cohort studies, from evidence of mechanisms, or from individual case studies (see preceding

[14]Our treatment of adjustment mirrors the discussion in the causal graph literature, where authors insist on proper adjustment for the computation of causal effects from observational data. Pearl (2000), e.g., devises his *do*-calculus precisely for this aim. Two graphical criteria (Pearl's front-door and back-door criteria) help the researcher decide whether the observational data at hand suffice tocalculate causal effects without supplementing interventional information. If so, the probability-theoretic instruments of front-door adjustment and back-door adjustment can be applied to calculate the desired measure (see, e.g., Shrier and Platt, 2008 for an application of Pearl's calculus in a medical setting). Despite the power of the calculus, the methodology naturally remains statistical (and Bayesian, so to speak), with all its shortcomings – e.g., the results must always be assessed relative to the prior choice of variables.

[15]Other kinds of biases, may also affect the reliability of the evidential reports. These regards the interest of researchers to get career advancements, or other kinds of incentives for scientific reputation. However, we follow recent suggestions (Bero and Grundy, 2016) to clearly distinguish these from the former kind, and we model here only the former one, also for the sake of keeping the model simple, and the case study manageable. However, our framework can be easily expanded to include several sources and types of biases.

[16]See also (Romero, 2016; Osimani and Landes, under review) for formal analyses of the role of bias and random error in hypothesis confirmation through replications.

[17]The statistical literature also offers tools for the detection of publication bias, e.g., funnel plots (Duval and Tweedie, 2000).These relate to the representativeness of the set of available studies with respect to the population characteristics.

[18]In individual case studies, time precedence is one of the main indicators of so called "token" causation, that is, actual cases of causation happening in individual patients. In this case, time precedence is not accompanied by statistical evidence of any kind, neither need it to come together with evidence about possible mechanisms of action, but rather stands alone.

[19]In the philosophical literature, the strength of a causal relationship has also been conceived as stability, invariance, insensitivity, or non-contingency. (Woodward, 2003) proposes a conceptualisation of strength of causal relationships in terms of invariance: Such relationships are not distorted nor disrupted – even under interventions in the system; they invariantly propagate causal influence. This way to conceive stability pertains to the relevance dimension (see Landes et al, 2018), that is to the context-sensitivity of causation, and has therefore implications for external validity (see also: Osimani, 2019). We will not treat these aspects in the present paper.

section). Longitudinal studies, may provide more data-points in time regarding the evolution of a phenomenon.[20]

A PROBABILISTIC INFERENCE MODEL: HYPOTHESIS UPDATING

We now show how one may model inferences from data to the causal hypothesis. Functional forms and concrete numbers are to be read as exemplary and can be found in the **Supplementary Material**.

Causal Variables

The Causal Hypothesis

The *prior* probability of the binary variable © is one's assessment of the probability that the drug causes the adverse effect after the hypothesis has been generated, without looking at any further evidence (only the evidence for generating the hypothesis may be taken into account here). The choice of a particular prior probability, $P(©)$, is hence *via* case-by-case reasoning.

Causal Indicators

The conditional probabilities of a causal indicator variable, given its parent variables, measure how much different types of evidence contribute towards confirmation of the causal hypothesis of interest.

The conditional probabilities relating © to the causal indicators are relatively stable across applications because they relate a theoretical entity, ©, to abstract indicators (Σ, Δ, M, T), see also Swaen and van Amelsvoort (2009) p. 272. Determining some of these conditional probabilities (see **Figure 4**) of indicator variables is straight-forward due to their entailment relationships. We have

$$\begin{aligned} P(M \mid ©) = P(T \mid ©) = P(\overline{\Delta} \mid \overline{©}) &= 1 \\ P(PD \mid DR) = P(DR \mid RoG) = P(T \mid M) &= 1, \end{aligned}$$

that is, the probability of there being a mechanism given that there is a causal relationship between $\mathcal{D}$ and E is one, just because $© \Rightarrow M$. The same holds for time course. Instead, the entailment relationship between Δ and © goes in the opposite direction: $\Delta \Rightarrow ©$. Hence, $\overline{©}$ entails $\overline{\Delta}$ and consequently the probability of $\overline{\Delta}$ given is $\overline{©}$ one.

Similarly, the probability of there being probabilistic dependence between $\mathcal{D}$ and E given that there is a dose-response relationship between them is also one. For the same reason, the probability of there being a dose-response relationship given that there is a high rate of growth is one. Finally, since M entails T, the probability of T given M is one.[21]

Statistical Indicators

One may assign the remaining conditional probabilities of the other indicators in Σ by first determining a finite number of curves relating dosage and adverse effect, which plausibly represent the possible dose-response curves. Next, one observes which of these curves are compatible with *PD*, *DR*, *RoG*, ©. Then, one assigns prior probabilities to these curves conditional on the causal hypothesis holding or not. This suffices to compute the remaining conditional probabilities. For example, $P(PD \mid ©, DR)$ is the probability of all curves when both *PD* and © hold but there is no dose-response relationship (*DR*), divided by the probability of all curves where © holds but not *DR*. So, $P(PD \mid © \overline{DR})$ is equal to:

$$\frac{P(\textit{curves exhibiting } PD, © \textit{ and } \overline{DR})}{P(\textit{curves exhibiting } © \textit{ and } \overline{DR})}$$

Difference Making

Since difference making is a very good indicator, we adopt "opinionated" conditional probabilities (reflecting tight relationships):

$$P(\Delta \mid ©) \approx 1 \quad \text{and} \quad P(\Delta \mid \overline{©}) = 0.$$

The first probability equals, in essence, 1 minus the probability of holistic causation. For the purposes of calculations in our case study we here set this value equal to 1.

Time

Time precedence is guaranteed either by there being a mechanism that leads from $\mathcal{D}$ to E, whether causal or not, or by there being a causal connection between $\mathcal{D}$ and E. This is because, if there exists a mechanism from $\mathcal{D}$ to E, then $\mathcal{D}$ must be prior to E.[22]

Also if $\mathcal{D}$ causes E, then $\mathcal{D}$ must be prior to E as well. So:

$$P(T \mid M, ©) = P(T \mid M, \overline{©}) = P(T \mid \overline{M}, ©) = 1.$$

The probability of there being time precedence is one, whenever either M or © hold.

In case M and © are false, we have no reasons to think that E is prior or posterior to $\mathcal{D}$. We are hence indifferent over whether there is time precedence or not. So,

$$P(T \mid \overline{M}, \overline{©}) = 0.5.$$

Mechanisms

If a drug causes a side effect, then this must occur *via* some mechanism, so $P(M|©)=1$. One's probability that there exists a physiological mechanism from drug to adverse effect, which is not causally responsible for the effect, is $P\left(M \mid \overline{©}\right)$. Since the probability of there being any physiological mechanism that goes

[20]Note that – under the assumption of a causal link between *D* and *E* – knowledge about the *T* indicator breaks the symmetric information provided by the *PD* indicator: E.g., if *D* is known tobe temporally prior to *E*, then the inference from (symmetric) probabilistic dependence between *D* and *E* to *E* causing *D* is not viable (*D* and *E* possessing a common cause is still a live option, though). Some philosophers of causation have made this explicit in their accounts of cause–effect relationships, see (Suppes, 1970).

[21]Since all causal indicator variables are binary, and probabilities sum to one, we only need to specify one half of all conditional probabilities in the Bayesian net.

[22]The fact that when $\mathcal{D}$ causes *E*, in an individual or at a population level, it must come before *E*, does not obviously entail that the drug $\mathcal{D}$ cannot also be taken after the event. We thank an anonymous reviewer for pointing this out to us.

from $\mathcal{D}$ to E, even if $\mathcal{D}$ does not cause E, is relatively high, see Howick (2011), we set this probability to 0.5.

Evidential Variables

Each study may yield evidence for (or against) any of our causal indicators. While experimental studies yield information about difference making in addition to probabilistic dependence and time (as well as, possibly, dose-response and rate of growth), observational studies may yield information about any one of the Σ indicators only (plus, possibly, information about time). Basic science studies or animal studies (or computational methods of various kinds) may deliver information about physiological mechanisms. See section on evidential mediators and reports (*Evidential Modulators: Study Design* and *Reports*).

We formalise the notion of incoming evidence as reports confirming or (dis-) confirming any of the indicators. These are represented in the Bayesian network as variables called *Rep* (for report). These report variables as well as the modulator variables (see *Statistical Evidence for the* Σ*-Indicators*, *Evidence of Difference-Making* below) are continuous variables, which, e.g., allow for the representation of an effect size, the duration of a study in days and the quality of randomisation.

Statistical Evidence for the Σ-Indicators

We assume that every observational study yields information about one Σ indicator only: i.e., each *Rep* node only has one Σ parent, graphically speaking. This parent is the strongest indicator one has evidence for. For instance, a multiple-exposure study delivering information about different effect sizes in the different arms with a steep rate of growth feeds into the *RoG* indicator only. Conversely, an observational study that delivers information about the outcome of exposed vs. non-exposed subjects only, with no graded arms differentiating among diverse dosages, will feed its evidence into the weakest Σ indicator only (*PD*).

For each observational study, the values of the following variables are pertinent for the report's conditional probability: adjustment for confounders *A*, sample size of the study *SS*, study duration *SD* and sponsorship bias *SB*.

The variables *A*, *SS* and *D* model how well a study tracks a Σ-indicator. The better the tracking the more informative a study is, the smaller the uncertainty, ceteris paribus. There may of course be other factors for a study ability to track a signal from nature that are outside of our model.

The presence of sponsorship bias instead, in the case of drug side-effects, is expected to lead to fewer reports of suspected adverse drug reactions and smaller effect sizes, i.e., side-effects tend to be concealed. The duration of a study is not a signal-tracking component in case a causal indicator does not hold, since whatever the length of the study, this will never detect a signal that nature does not send.

Figures 5 and **6** (for positively and negatively instantiated *PD*, respectively) compactly illustrate these shifting tendencies when these dimensions interact. The graphs show for a (non-)significant effect size, $ES \in \{0, 1\}$, how the conditional probability of a report changes (in tendency) when the sponsorship bias variable *SB* and the signal-tracking (as a composite variable) change. Case

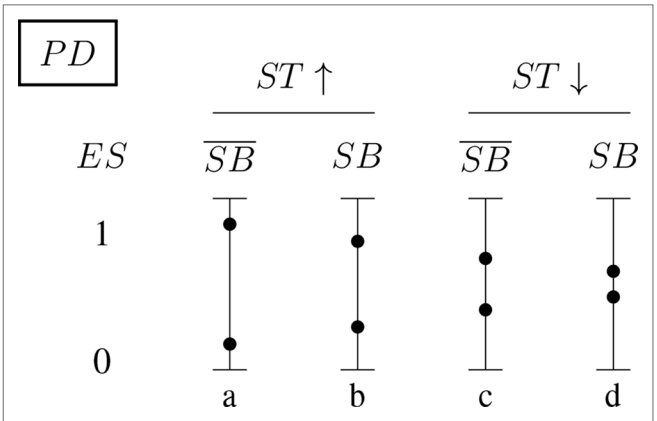

FIGURE 5 | Conditional probabilities of a report ($ES \in \{0, 1\}$) for the positively instantiated *PD* (probabilistic dependence) indicator: The four bars illustrate the shift in tendency from scenario (a), with better signal-tracking and no sponsorship bias, to scenario (d), with worse signal-tracking and sponsorship bias. The probability of a report of significant effect size, $ES = 1$, decreases from left to right.

(a) represents a better signal-tracking and no sponsorship bias, case (d) represents a worse signal-tracking and the presence of sponsorship bias, that is, the tendency to hide harmful effects. For example, for positively instantiated *PD* (**Figure 5**), adding the presence of sponsorship bias compresses the range. Worsening the signal-tracking (e.g., due to reduced sample size) also has this compression effect. Consider the case of a study which reports no adverse effect: if it is good at signal-tracking and has no sponsorship bias, then the probability of reporting such a null result is low, but it increases when sponsorship bias is present.

Evidence of Difference-Making

RCTs inform us about the difference-making indicator of causation and whether there is time precedence. For each study,

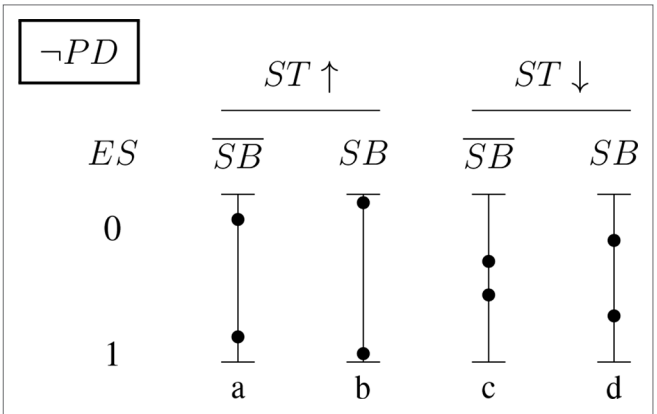

FIGURE 6 | Conditional probabilities of a report ($ES \in \{0, 1\}$) for the negatively instantiated *PD* (probabilistic dependence) indicator: As above, the four bars illustrate the shift in tendency from scenario (a), with better signal-tracking and no sponsoring bias, to scenario (d), with worse signal-tracking and sponsorship bias. In this case, the probability of a report of significant effect size, $ES = 1$, increases with worsening signal-tracking, but decreases when sponsorship bias is accounted for.

the report's conditional probability depends on the variables we used for statistical evidence, (adjustment, sample size, duration, sponsorship bias: *A*, *SS*, *D*, *SB*), plus: blinding *B*, randomisation *R* and placebo *Pl*. Ceteris paribus, the better blinding, randomisation and placebo implementation the better a study is at tracking the signal, or, in case no signal needs to be detected, the more it reduces the chances of false positives.

Assessment of Modulators

The assessment of the modulators *SS*, *D* is achieved by reading off study characteristics of published reports. There is hence no uncertainty about these variables. As a result, there is no need to explicitly represent these modulators as variables in the Bayesian network.

The other modulators may be assessed by the application of quality assessment tools (QATs). In case there is uncertainty about a particular modulator applying to a study, which may be due to disagreement between different QATs Stegenga (2014) or to lack of available data, this modulator is represented by a variable *V*. The uncertainty over *V* then leads to what Bayesian statisticians call a *hierarchical model*. Instead, for a Bayesian epistemologists the modulator variable *V* is a variable like any other and she is hence prepared to assign (conditional) probabilities to it. Technically, one specifies an unconditional probability distribution over *V* reflecting this uncertainty. In the DAG one adds an arrow starting at *V* which points to the report variable. The conditional probabilities of the report variable is then specified with respect to all the possible values of all its parents (including *V*).

If an (a group of) author(s) is responsible for multiple reports which may affected from sponsorship bias, then one creates only a single variable *V* for this (group of) author(s) which modulates all these studies. This construction allows one to reason about the sponsorship bias of the (group of) author(s) from data.

Mechanistic Evidence

Studies at the genetic, molecular, or cell level are often considered to provide evidence about the mechanisms that underpin the putative phenotypic causal relation. This observation motivates our choice of introducing a variable M_i for every mechanism for which there is evidence. The M_i come as hypotheses about mechanisms between $\mathcal{D}$ and *E*. Each mechanism M_i may be broken down into further bits of the mechanism; denoted here by $\mu_{i,k}$. Concrete data about these bits is denoted by $Rep_{\mu_{ik}}$, see **Figure 7** for an illustration.

As for the reports feeding into the Σ set or to Δ, also the $Rep_{\mu_{i,k}}$ reports might be modulated by evidential modulators. However, evidential modulators are of a different nature here and deserve a separate treatment. Hence, in order to keep this paper self-contained and not to complicate calculations for the case study excessively, we do not model here evidential modulators for evidence of mechanisms.

Evidence of the Temporal Structure

Evidence of the temporal structure comes from RCTs, and also cohort studies which can reduce the suspicion of reverse causation, but not other confounders. Modulo other confounders, a cohort study reporting an observed effect provides evidence for a statistical correlation and the temporal structure, at the same time.

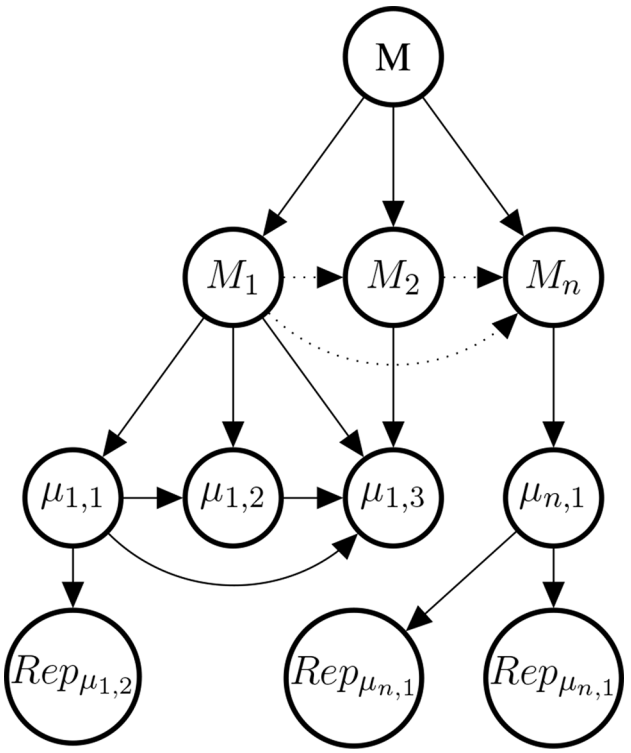

FIGURE 7 | Illustrative example of the mechanism part of the Bayes net. The graph shows the existential claim *M* (mechanistic knowledge) and its relationship with hypothetical, alternative mechanisms $M_1, M_2, \ldots, M_n$, their constitutive sub-mechanisms ($\mu_{i,k}$) and concrete evidence $\left(Rep_{\mu_{i,k}}\right)$. Dotted edges are present, if and only if two M_i share parts of their mechanisms. Sub-mechanisms nodes ($\mu_{i,k}$) without children are to be read as hypothesised sub-mechanisms for which no evidence is available. Every sub-mechanism may have multiple evidence reports as children which may represent basic science findings in different species or cell cultures.

APPLICATION OF THE MODEL: DOES PARACETAMOL CAUSE ASTHMA?

In the following, we apply our framework to a case study: the debated causal association between paracetamol and asthma. The debate is not settled yet (Heintze and Petersen, 2013; Henderson and Shaheen, 2013; Martinez-Gimeno and García-Marcos, 2013)[23] and evidence concerning this hypothesis is by now considerably vast and varied. For simplicity, we will here consider only exemplary studies in the entire body of now available evidence, and simulate on the basis of these studies, how hypothesis updating could be modelled in our framework. We specified the causal variables and their conditional (in-)dependencies in *Theoretical Entities*. The report variables for statistical and difference-making evidence and their conditional (in-)dependencies are described in

[23]An older reference which also gives an assessment of © along the Bradford Hill Criteria is (Farquhar et al., 2009, Page 39).

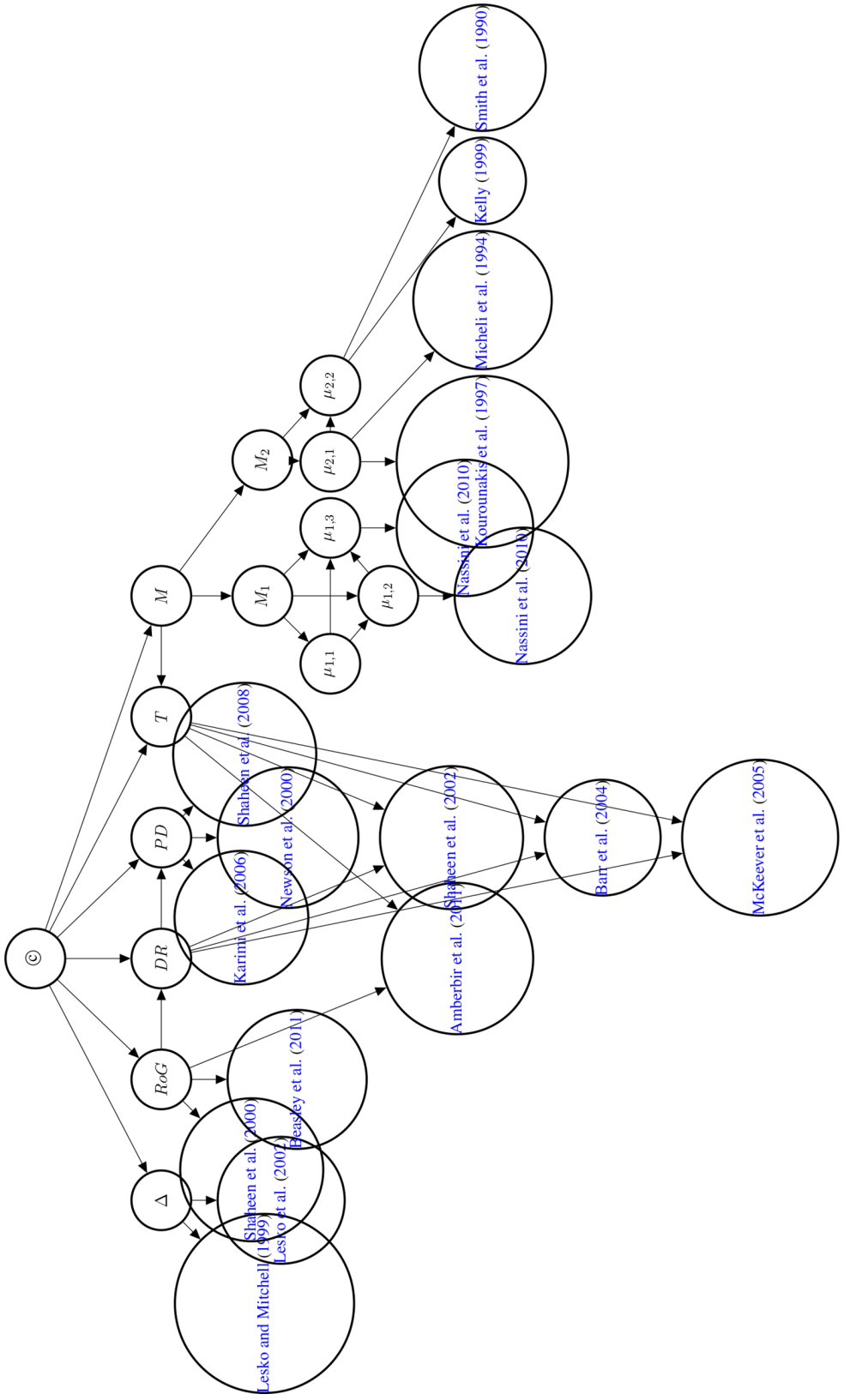

FIGURE 8 | Directed acyclic graph of the Bayesian network used to compute the posterior probability of © (Hypothesis of Causation). Evidential modulators are not shown.

Evidential Variables; their conditional probabilities are specified in Section 2 of the **Supplementary Material**. How to set up the mechanistic part of the model is explained in Section 4 of the **Supplementary Material**.

Although, evidence ought to be considered always with respect to a given population of interest; we do not make any such distinction here for the sake of a compact presentation.

We here present summaries of reported results, for none of which we claim any credit.

Hypothesis Generation

The hypothesis of a possible causal association between paracetamol intake and asthma first emerged following the observation that the "asthma epidemic" in the western population followed the same time trend of increase in paracetamol consumption.

The data on which this observation was based initially came from a study by Varner et colleagues (Varner et al., 1998). The study aimed at explaining this epidemic through the reduction of aspirin use in the same period, due to the protective properties of aspirin against asthma in virtue of its anti-inflammatory effects. Aspirin prescription declined because the drug was discovered to be associated with Reye's syndrome (Varner et al., 1998).

The hypothesis that asthma epidemic could be explained by the drop of aspirin prescription, was however undermined by simply considering that, if it were true, then one should have observed an equal prevalence of asthma *before* aspirin was introduced into the market, and a decrease after its introduction (Shaheen et al., 2000). Since in the same study the data-points showed a coincidence in time trends not only between asthma increase and aspirin decline, but also between increase of paracetamol sales and of asthma prevalence, this led researchers to investigate the causal hypothesis that paracetamol causes asthma; see Henderson and Shaheen (2013); Osimani (2014) for more details.

However, at the time in which this hypothesis was generated, there was little belief that the household paracetamol may be causing asthma, because of a general assumption of innocuousness. Experts at the time of hypothesis generation hence had a low prior belief in © begin true. Since we do not have access to a time travel machine, we exemplary consider three plausible values of the prior probability $P(©)$: 0.01,0.005 and 0.001 for illustrative purposes.

Statistical Evidence

The statistical and mechanistic evidence presented next is a small part of all the available evidence concerning the debated causal connection between paracetamol and asthma. Studies were selected to demonstrate the workings of the model and its versatility: some of these studies are shown below, other ones are presented in the **Supplementary Material**; exhaustiveness and representativeness were not part of our study selection procedure.

To simplify exposition we here model a state in which there is no uncertainty about the modulators applying to evidential reports, that is, one is sure whether a study is properly adjusted, blinded and so on. Furthermore, we limit ourselves here to binary effect size variables $ES \in \{0, 1\}$ and discrete modulator variables in $\{0, 0.5, 1\}$ about which we are certain. In the **Supplementary Material** (*Methodology*), we explain how to model uncertainty about the value of modulators variables *via* Bayesian hierarchical modelling.

Lesko and Mitchell (1999) reports a practitioner-based, double-blind, clinical trial, with random assignment of paracetamol and Ibuprofen to 27,065 children, without placebo, and with a 4-week follow-up period. The aim of the study was to investigate the safety of ibuprofen, rather than paracetamol. Relevant outcomes were hospitalisation for asthma/bronchiolitis; the relative risk for ibuprofen, compared with paracetamol was 0.9 (95% CI, 0.5–1.4). Since the confidence interval for the relative risk contains 1, there is no evidence of either of the two being more or less harmful to children. With regard to a possible sponsorship bias, this study was reported to be supported by McNeil Consumer Products Company, Fort Washington, Pennsylvania.[24] Since the study was run without placebo and for a relatively short period, the probability of observing a null effect, as in this case, is relatively high. Furthermore, the observed null effect may be due to a) neither the drug being harmful or b) both drugs being harmful. However, this latter possibility is excluded, through implicit comparison to the base-rate incidence in the overall population. Hence, we consider this study, notwithstanding its lack of placebo, to feed into the Δ indicator. In order to update our hypothesis on this evidence ($ES = 0$), we need to fully specify all conditional probabilities of observing it, when the pertinent indicator(s) hold [or not] given the evidential modulators. We assess the modulators for this study as follows: $A = 0.5$, $SS = 1$, $D = 0$, $SB = 1$, $B = 1$, $R = 1$, $Pl = 0.5$. We use $\vec{x}$ to denote the values of the pertinent modulators here and in the following formulae.[25]

$$
\begin{aligned}
P(ES=0\,|\,\vec{x},\Delta) &= 1-P(ES=1\,|\,\vec{x},\Delta)\\
&= 1-(1-\frac{SB}{10})\cdot(1-0.5\cdot\|1-avg(A,SS,D,B,R,Pl)\|)\\
&= 1-0.9\cdot\left(0.5+\frac{0.5+1+0+0+1+0.5}{12}\right)=\frac{13}{40}\\
P(ES=0\,|\,\vec{x},\bar{\Delta}) &= 1-P(ES=1\,|\,\vec{x},\bar{\Delta})\\
&= 1-(1-\frac{SB}{10})\cdot\left(0.5\cdot\|1-avg(A,SS,B,R,Pl)\|\right)\\
&= 1-0.9\cdot\left(0.5\cdot\left(1-\frac{0.5+1+1+1+0.5}{5}\right)\right)=\frac{91}{100}.
\end{aligned}
$$

This formula captures the idea that, if a study is good at tracking the signal, then the probability of observing the effect, given that the related statistical indicator holds, tends to 1. Instead, the worse the study is, the smaller the probability becomes. See Section 2 of the **Supplementary Material** for further details.

Shaheen et al. (2002) reports a population based longitudinal study (Avon study). Observations are reported at different times,

[24]McNeil Consumer Products Company and Whitehall-Robins Healthcare (Madison, NJ). The Slone Epidemiology Unit has received or is currently receiving research support from the US Food and Drug Administration, the National Institutes of Health, and a number of pharmaceutical companies.

[25]All conditional probabilities of evidential variables are here stated prior to normalization (see the **Supplementary Material (Section 2)** for rationales for normalization).

for a minimum of 9,400 patients: pregnant women and their babies of up to 42 months. After controlling for potential confounders, frequent paracetamol use in late pregnancy (20-32 weeks), but not in early pregnancy (< 18-20 weeks), was associated with an increased risk of wheezing in the offspring at 30-42 months (adjusted odds ratio (OR) compared with no use 2.10 (95% CI 1.30 to 3.41); p = 0.003), particularly if wheezing started before 6 months (OR 2.34 (95% CI 1.24 to 4.40); p = 0.008). Assuming a causal relation, only about 1% of wheezing at 30-42 months was attributable to this exposure. Two authors of this study (SOS and RBN) report funding from the UK Department of Health. Core funding for the long term follow up of the cohort came from the Medical Research Council, the Wellcome Trust, the UK Department of Health, the Department of the Environment, DfEE, the National Institutes of Health, a variety of medical research charities and commercial sponsors, including Stirling-Winthrop who enabled the original collection of data on paracetamol use. We model this as evidence pertaining to *DR* and *T* (since only two different non-zero dosages – never, some days, most days– were reported). For the modulators we have $SS = 1$, $D = 1$, $SB = 0$, $A = 1$ and thus

$$P(ES = 1 \mid \vec{x}, DR, T) = 0.5 + \frac{avg(A, SS, D)}{2} = 1$$
$$P(ES = 1 \mid \vec{x}, \overline{DR, T}) = 0.5 \cdot \| 1 - avg(A, SS) \| = 0.$$

Mechanistic Evidence

To focus the exposition, we only consider two possible mechanisms (M_1 and M_2) by which paracetamol may cause asthma.

M_1: Paracetamol is metabolised to NAPQI (N-acetyl-p-benzoquinone imine) ($\mu_{1,1}$), NAPQI stimulates transient receptor potential ankyrin-1 (TRPA1) ($\mu_{1,2}$) [reported in Nassini et al. (2010)] and TRPA1 causes airway neurogenic inflammation ($\mu_{1,3}$) [reported in Nassini et al. (2010)].

M_2: Paracetamol depletes Gluthatione ($\mu_{2,1}$) [reported in Micheli et al. (1994); Kourounakis et al. (1997)], low levels of Gluthatione cause oxidative stress hyperresponsiveness in the airways ($\mu_{2,2}$) [reported in Smith et al. (1990); Kelly (1999)].

We set the conditional probabilities of a mechanism given M to:

$$P(M_1 \mid M) = 0.7 \qquad P(M_1 \mid \bar{M}) = 0$$
$$P(M_2 \mid M) = 0.8 \qquad P(M_2 \mid \bar{M}) = 0.$$

We assessed M_1 and M_2 to be likely, if M holds; M_1 was assessed to be the more likely of the two. If M does not hold, then all M_i have to fail to hold and are hence assigned zero probability.

We now turn to setting conditional probabilities of the $\mu_{1,k}$ given M_1 and given $\overline{M_1}$ First, recall that M_i entails $\mu_{i,k}$ and hence

$$P(\mu_{1,1} \mid M_1) = 1$$
$$P(\mu_{1,2} \mid M_1, \mu_{1,1}) = 1$$
$$P(\mu_{1,3} \mid M_1, \mu_{1,1}, \mu_{1,2}) = 1.$$

$\mu_{1,1}$, $\mu_{1,2}$, $\mu_{1,3}$ and $\overline{M_1}$ are, when taken together, logically inconsistent. So,

$$P(\mu_{1,3} \mid \overline{M_1}, \mu_{1,1}, \mu_{1,2}) = 0.$$

If M_1 fails to hold, then we are indifferent about $\mu_{1,3}$ and $\mu_{1,2}$ – independently of $\mu_{1,1}$ (respectively $\mu_{1,2}$).

$$P(\mu_{1,2} \mid \overline{M_1}, \mu_{1,1}) = P(\mu_{1,2} \mid \overline{M_1}, \overline{\mu_{1,1}}) = 0.5$$

$$P(\mu_{1,3} \mid \overline{M_1}, \overline{\mu_{1,1}}, \mu_{1,2}) = P(\mu_{1,3} \mid \overline{M_1}, \mu_{1,1}, \overline{\mu_{1,2}}) = P(\mu_{1,3} \mid \overline{M_1}, \overline{\mu_{1,1}}, \overline{\mu_{1,2}}) = 0.5.$$

In general, almost all effective drugs have toxic metabolites. We here take it as established that paracetamol is metabolised to NAPQI (independently of whether M_1 holds or not) and hence put

$$P(\mu_{1,1} \mid \overline{M_1}) = 1.$$

Conditional probabilities of considered evidence reports in (Nassini et al., 2010) for M_1 are set to:

$$P(Rep1_{\mu_{1,2}} \mid \mu_{1,2}) = 0.91 \qquad P(Rep1_{\mu_{1,2}} \mid \overline{\mu_{1,2}}) = 0.09$$
$$P(Rep2_{\mu_{1,3}} \mid \mu_{1,3}) = 0.91 \qquad P(Rep2_{\mu_{1,3}} \mid \overline{\mu_{1,3}}) = 0.09$$

We take the quotient $P(Rep_{\mu_{1,2}} \mid \mu_{1,2}) / P(Rep_{\mu_{1,2}} \mid \overline{\mu_{1,2}})$ to be a measure of the strength of evidence in accordance with the literature on Bayes factors. It expresses how much more (or less) likely the received evidence is under μ than under $\bar{\mu}$ A Bayes factor of 91/9 ≈ 10 was chosen to model confident claims in the primary literature, while a Bayes factor of 75/25 = 3 was adopted for cautious claims. Conditional probabilities of considered evidence reports for M_2:

$$P(\mu_{2,1} \mid M_2) = 1 \qquad P(\mu_{2,1} \mid \overline{M_2}) = 0.01$$
$$P(\mu_{2,2} \mid M_2, \mu_{2,1}) = 1$$
$$P(\mu_{2,2} \mid \overline{M_2}, \mu_{2,1}) = 0 \qquad P(\mu_{2,2} \mid \overline{M_2}, \overline{\mu_{2,1}}) = 0.5.$$

$P(\mu_{2,2} \mid \overline{M_2}, \mu_{2,1})$ is zero for the same reasons as $P(\mu_{1,3} \mid \overline{M_1}, \mu_{1,1}, \mu_{1,2})$ is equal to zero. Conditional probabilities of considered evidence reports (Smith et al., 1990; Micheli et al., 1994; Kourounakis et al., 1997; Kelly, 1999) are set to

$$P(Rep1_{\mu_{2,1}} \mid \mu_{2,1}) = 0.91 \qquad P(Rep1_{\mu_{2,1}} \mid \overline{\mu_{2,1}}) = 0.09$$
$$P(Rep2_{\mu_{2,1}} \mid \mu_{2,1}) = 0.91 \qquad P(Rep2_{\mu_{2,1}} \mid \overline{\mu_{2,1}}) = 0.09$$
$$P(Rep1_{\mu_{2,2}} \mid \mu_{2,2}) = 0.75 \qquad P(Rep1_{\mu_{2,2}} \mid \overline{\mu_{2,2}}) = 0.25$$
$$P(Rep2_{\mu_{2,2}} \mid \mu_{2,2}) = 0.91 \qquad P(Rep2_{\mu_{2,2}} \mid \overline{\mu_{2,2}}) = 0.09.$$

TABLE 2 | Posterior Probability of © (Hypothesis of Causation) with accumulating evidence. Every row indicates the probability of © given the body of evidence up to and including this row. Nassini et al. (2010) reports evidence for two different nodes in the Bayesian network and is hence listed twice here.

No Evidence	Prior Probability of ©	0.0100	0.0050	0.0010
Mechanistic Evidence	Smith et al. (1990)	0.0175	0.0088	0.0018
	Micheli et al. (1994)	0.0193	0.0097	0.0019
	Kourounakis et al. (1997)	0.0195	0.0098	0.0020
	Kelly (1999)	0.0196	0.0098	0.0020
	Nassini et al. (2010)a	0.0196	0.0099	0.0020
	Nassini et al. (2010)b	0.0198	0.0099	0.0020
Statistical Evidence discussed above	Lesko and Mitchell (1999)	0.0072	0.0036	0.0007
	Shaheen et al. (2002)	0.1534	0.0827	0.0176
Statistical Evidence discussed in the **Supplementary Material**	Shaheen et al. (2000)	0.2238	0.1254	0.0278
	Newson et al. (2000)	0.0997	0.0522	0.0109
	Lesko et al. (2002)	0.3686	0.2250	0.0547
	Barr et al. (2004)	0.6397	0.4690	0.1496
	McKeever et al. (2005)	0.6445	0.4742	0.1523
	Karimi et al. (2006)	0.6446	0.4743	0.1523
	Shaheen et al. (2008)	0.6446	0.4743	0.1523
	Amberbir et al. (2011)	0.7055	0.5437	0.1918
	Beasley et al. (2011)	0.7160	0.5564	0.1999
All evidence discussed here	Posterior Probability of ©	0.7160	0.5564	0.1999

The first three reports are assessed as confident claims, the fourth claim as cautious. The graph of the Bayesian network is displayed in **Figure 8**.

Posterior Probability of ©

The body of evidence incorporated here was assembled to show the versatility of the E-Synthesis framework and by no way represents a systematic review of the available evidence. The posterior probability is thus best understood as an illustration of how the framework computes the posterior without attaching too much weight to the actual computed value. We present the posteriors for the three different priors (0.01, 0.005, 0.001) as an example of a sensitivity analysis investigating the output (posterior probability of ©) on the input parameters.

We computed the posterior probability of © using the Bayes Net Toolbox in Matlab R2012a and report the posterior probabilities in **Table 2**.[26] Formally, computing the posterior probability of © is a Bayesian network inference problem which can be solved by repeated applications of the Chain Rule and Bayes' Theorem, see (Neapolitan, 2003).

DISCUSSION

Learning Probabilities From Data

The (conditional) probabilities introduced above are a general example which may reflect the judgements of expert opinions. In concrete applications, these probabilities can and should heavily draw on real-world evidence and data. There is some work on how to determine such probabilities in this way, as we now briefly outline.

De Pretis and Osimani (2019) suggest an approach to compute the following:

$$P(DR \mid \mathcal{E}_{DR})$$

i.e., the probability of *DR*, given the available dose-response evidence and its related modulators. The MCP-Mod algorithm (see Bretz et al., 2005) and similar Bayesian approaches (Shao and Shapiro, 2019) have been proposed to tackle the problem of dose-finding in pre-clinical studies aiming to determine the (optimal) *efficacy* of a drug. These algorithms consider a finite number of dose-response curves, which is similar to our approach (**Figure 3**). Unfortunately, the approach of De Pretis and Osimani (2019) does not yet involve the role of modulators, that is, they do not compute $P(DR \mid \mathcal{E}_{DR}, \vec{x})$. We hence cannot simply incorporate their results.

Stewart et al. (2015) presents a Bayesian network approach to judge the quality of studies within the GRADE framework, see Alonso-Coello et al. (2016). "The approach also lends itself to automation, where nodes can be parameterised either using data mining software." (Stewart et al., 2015, Page 9).

Ryan et al. (2013) determines a conditional probability of © given a statistical association. Unfortunately, we cannot use this probability here because the conditional probabilities of *PD* we require here are also conditionalised on the (non-) existence of a dose-response.

Evidence Synthesis in Context

Theoretically and computationally, E-Synthesis exploits coherence of partly or fully independent evidence converging towards the hypothesis (or of conflicting evidence with respect to it), in order to update its posterior probability. Propagation of probabilities hence work in a totally different sense than for causal DAGs (Pearl, 2000; Spirtes et al., 2000). Probabilities reflect here epistemic uncertainty and, loci of uncertainty are made transparent in terms of articulated (conditional) probabilities, as well as graphically traceable in terms of a DAG.

With respect to other frameworks for evidence synthesis (Greenhalgh et al., 2011; Kastner et al., 2012; Warren et al.,

[26]The underlying DAG is displayed in **Figure 8**. The Matlab R2012a .m-file used to compute the poterior probability can be found in the **Supplementary Material**.

2012; van den Berg et al., 2013; Kastner et al., 2016; Tricco et al., 2016a; Tricco et al., 2016b; Shinkins et al., 2017), our Bayesian model has the unique feature of grounding its inferential machinery on a consolidated theory of hypothesis confirmation (Bayesian epistemology), and in allowing any data from the most heterogeneous sources (cell-data, clinical trials, epidemiological studies), and methods (e.g. frequentist hypothesis testing, Bayesian adaptive trials, etc.) to be quantitatively integrated into the same inferential framework. E-Synthesis is thus at the same time highly flexible concerning the allowed input, while at the same time relying on a consistent computational system, philosophically and statistically grounded.

By introducing evidential modulators, and thereby breaking up the different dimensions of evidence (strength, relevance, reliability) E-Synthesis allows them to be explicitly tracked in the body of evidence. This makes it possible to parcel out the strength of evidence from the method with which it was obtained.[27] With this, E-Synthesis provides a higher order perspective on evidential support by effectively embedding these various epistemic dimensions in a concrete topology.

CONCLUSIONS

This paper focuses on inference *within one model*, rooting in *one hypothesis*, but E-Synthesis allows for going beyond the network limits and for embedding it in an even larger network to trace the hypothesis relation with other potentially concurring hypotheses. The mechanics of Bayesian epistemology are flexible enough to permit such an augmentation for the purposes of tracing further inference patterns.[28] For simplicity's sake we have not presented in the current paper dimensions of evidence relating to external validity and extrapolation; however the framework itself already incorporates also this sort of evidential modulators [see (Landes et al., 2018)]. We will illustrate the functioning of these modulators in a separate paper. Further limitations are that all causal indicator variables and © (in particular, there is no way to express and reason about the strength of causation in ©), conditional probabilities were not set *via* expert elicitation and the general formulae for the conditional probabilities of evidential variables (Section 2 in the **Supplementary Materials**) for illustrative purposes.

Future work may take a number of directions such as developing scoring methods learned from data (*Discussion*) and/or based on expert opinions, applications to further case studies to demonstrate the versatility of the framework, analysis and incorporation of further evidential modulators (those mentioned at the end of *Evidential Modulators: Study Design* as well as modulators of external validity), analysis and incorporation of further biases (a catalogue of biases is currently developed at the Oxford Centre for EBM), comparing E-Synthesis to other frameworks of causal assessment *via* applications to the same case study (Abdin et al., 2019) and formally modelling and incorporating (spontaneous) case reports, evidence obtained *via* text mining and/or data base search into the framework. Finally, applying non-binary report variables to capture odds ratios, relative risks and/or confidence intervals are subject to future study.

[27]Osimani and Landes investigate in (Osimani and Landes, under review) various concepts of reliability involved in such considerations.

[28]Even if one is unenthusiastic with respect to the Bayesian approach: s/he can take our framework as a way to consistently structure the aggregation of evidence – as it is implicitly carried out in systematic reviews of quantitative and qualitative studies.

AUTHOR CONTRIBUTIONS

BO developed the idea of implementing Bradford-Hill criteria into an epistemic Bayesian net, identified relevant issues in the epistemology of causation as well as in the current debate on causal and statistical inference in medicine and especially in pharmacosurveillance. JL proved the mathematical soundness of the evidence aggregation tool and helped with the analysis of emerging statistical and methodological issues. FP reviewed the paper from a mathematical point of view and contributed to address *E-Synthesis* to a pharmacovigilance perspective.

FUNDING

The research for this paper was funded by the European Research Council (grant 639276), the Marche Polytechnic University (Italy) and the Munich Center for Mathematical Philosophy (MCMP, Germany). JL and FP worked at the paper as 100% research fellows within the project, whereas BO is the project PI. For the final phase of writing this manuscript JL gratefully acknowledges funding from the German Research Foundation for the grant agreements LA 4093/2-1 (Evidence and Objective Bayesian Epistemology) and LA 4093/3-1 (Foundations, Applications & Theory of Inductive Logic). JL gratefully acknowledges funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 432308570 and 405961989.

ACKNOWLEDGMENTS

We would like to thank the two reviewers, Jeffrey K. Aronson, Daniel J. Auker-Howlett, Ralph Edwards, Niklas Norén, David Teira, Adam La Caze, Ulrich Mannsman, Stephan Lehr, Patrick Maison, Mehdi Benkebil, Malak Abou Taam, Maria Luisa Casini, and Martin Posch for helpful comments. We also want to thank audiences in Bonn, Bologna, Bristol, Cologne, Copenhagen, Exeter, Edinburgh, London, Munich, Sidney, Tilburg, Turin, Vienna, Canterbury, Oxford, Paris, Uppsala, Hanover, where parts of this work were presented.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2019.01317/full#supplementary-material

REFERENCES

Abdin, A. Y., Auker-Howlett, D., Landes, J., Mulla, G., Jacob, C., and Osimani, B. (2019). Reviewing the mechanistic evidence assessors E-Synthesis and EBM+: a case study of amoxicillin and drug reaction with Eosinophilia and systemic symptoms (DRESS). *Curr. Pharm. Des*. 25 (16), 1866–1880. doi: 10.2174/1381612825666190628160603

Adami, H. O., Berry, S. C. L., Breckenridge, C. B., Smith, L. L., Swenberg, J. A., Trichopoulos, D., et al. (2011). Toxicology and epidemiology: improving the science with a framework for combining toxicological and epidemiological evidence to establish causal inference. *Toxicol. Sci.* 122, 223–234. doi: 10.1093/toxsci/kfr113

Alonso-Coello, P., Schünemann, H. J., Moberg, J., Brignardello-Petersen, R., Akl, E. A., Davoli, M., et al. (2016). GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ* 353. doi: 10.1136/bmj.i2016

Amberbir, A., Medhin, G., Alem, A., Britton, J., Davey, G., and Venn, A. (2011). The role of acetaminophen and geohelminth infection on the incidence of wheeze and eczema. *Am. J. Respir. Crit. Care Med*. 183, 165–170. doi: 10.1164/rccm.201006-0989OC

Ankley, G. T., Bennett, R. S., Erickson, R. J., Hoff, D. J., Hornung, M. W., Johnson, R. D., et al. (2010). Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. *Environ. Toxicol. Chem.* 29, 730–741. doi: 10.1002/etc.34

Barr, R. G., Wentowski, C. C., Curhan, G. C., Somers, S. C., Stampfer, M. J., Schwartz, J., et al. (2004). Prospective study of acetaminophen use and newly diagnosed asthma among women. *Am. J. Respir. Crit. Care Med.* 169, 836–841. doi: 10.1164/rccm.200304-596OC

Beasley, C. M., Dellva, M. A., Tamura, R. N., Morgenstern, H., Glazer, W. M., Ferguson, K., et al. (1999). Randomised double-blind comparison of the incidence of tardive dyskinesia in patients with schizophrenia during long-term treatment with olanzapine or haloperidol. *Br. J. Psychiatry* 174, 23–30. doi: 10.1192/bjp.174.1.23

Beasley, R. W., Clayton, T. O., Crane, J., Lai, C. K. W., Montefort, S. R., von Mutius, E., et al. (2011). Acetaminophen use and risk of asthma, rhinoconjunctivitis, and eczema in adolescents. *Am. J. Respir. Crit. Care Med.* 183, 171–178. doi: 10.1164/rccm.201005-0757OC

Bero, L. A., and Grundy, Q. (2016). Why having a (nonfinancial) interest is not a conflict of interest. *PloS Biol.* 14, 1–8. doi: 10.1371/journal.pbio.2001221

Bogen, J., and Woodward, J. (1988). Saving the phenomena. *Philos. Rev.* 97, 303–352. doi: 10.2307/2185445

Bovens, L., and Hartmann, S. (2003). *Bayesian Epistemology* (Oxford: Oxford University Press).

Bretz, F., Pinheiro, J. C., and Branson, M. (2005). Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics* 61, 738–748. doi: 10.1111/j.1541-0420.2005.00344.x

Cartwright, N. (2001). What is wrong with bayes nets? *Monist* 84, 242–264. doi: 10.5840/monist20018429

Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics* (Cambridge: Cambridge University Press).

Caster, O., Sandberg, L., Bergvall, T., Watson, S., and Norén, G. N. (2017). vigiRank for statistical signal detection in pharmacovigilance: first results from prospective real-world use. *Pharmacoepidemiol Drug Saf.* 26, 1006–1010. doi: 10.1002/pds.4247

Cooper, N., Coyle, D., Abrams, K., Mugford, M., and Sutton, A. (2005). Use of evidence in decision models: an appraisal of health technology assessments in the UK since 1997. *J. Health Serv. Res. Policy* 10, 245–250. doi: 10.1258/135581905774414187

Dawid, A. P., Musio, M., and Fienberg, S. E. (2016). From statistical evidence to evidence of causality. *Bayesian Anal.* 11, 725–752. doi: 10.1214/15-ba968

Dawid, A. P. (2017). On individual risk. *Synthese* 194, 3445–3474. doi: 10.1007/s11229-015-0953-4

De Pretis, F., and Osimani, B. (2019). New insights in computational methods for pharmacovigilance: E-synthesis, a bayesian framework for causal assessment. *Int. J. Environ. Res. Public Health* 16 (12), 2221. doi: 10.3390/ijerph16122221

Duval, S., and Tweedie, R. (2000). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 56, 455–463. doi: 10.1111/j.0006-341X.2000.00455.x

Edwards, I. R., and Aronson, J. K. (2000). Adverse drug reactions: definitions, diagnosis, and management. *Lancet* 356, 1255–1259. doi: 10.1016/S0140-6736(00)02799-9

Egger, M., Schneider, M., and Smith, G. D. (1998). Spurious precision? Meta-analysis of observational studies. *BMJ* 316, 140–144. doi: 10.1136/bmj.316.7125.140

Ehmann, F., Papaluca-Amati, M., Salmonson, T., Posch, M., Vamvakas, S., Hemmings, R., et al. (2013). Gatekeepers and enablers: how drug regulators respond to a challenging and changing environment by moving toward a proactive attitude. *Clin. Pharmacol. Ther.* 93, 425–432. doi: 10.1038/clpt.2013.14

European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC) (2007). Intelligent Testing Strategies in Ecotoxicology: Mode of Action Approach for Specifically Acting Chemicals. Tech. rep. http://www.ecetoc.org/wp-content/uploads/2014/08/ECETOC-TR-102.pdf.

European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC). (2009). Framework for the Integration of Human and Animal Data in Chemical Risk Assessment. Tech. rep. http://www.ecetoc.org/uploads/Publications/documents/TR%20104.pdf.

European Commission (2008). Strengthening pharmacovigilance to reduce adverse effects of medicines. http://europa.eu/rapid/press-release_MEMO-08-782_de.htm?locale=en.

Farquhar, H., Stewart, A., Mitchell, E., Crane, J., Eyers, S., Weatherall, M., et al. (2009). The role of paracetamol in the pathogenesis of asthma. *Clin. Exp. Allergy* 40, 32–41. doi: 10.1111/j.1365-2222.2009.03378.x

Fedak, K. M., Bernal, A., Capshaw, Z. A., and Gross, S. (2015). Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. *Emerging Themes Epidemiol.* 12, 14. doi: 10.1186/s12982-015-0037-4

Fuller, J. (2019). The confounding question of confounding causes in randomized trials. *Br. J. Philosophy Sci.* doi: 10.1093/bjps/axx015

Gelman, A., and Hennig, C. (2017). Beyond subjective and objective in statistics. *J. R. Stat. Soc. Ser. (Statistics In Society)* 180, 967–1033. doi: 10.1111/rssa.12276

Geneletti, S., Gallo, V., Porta, M., Khoury, M. J., and Vineis, P. (2011). Assessing causal relationships in genomics: From bradford-hill criteria to complex gene-environment interactions and directed acyclic graphs. *Emerging Themes Epidemiol.* 8, 5. doi: 10.1186/1742-7622-8-5

Greenhalgh, T., Wong, G., Westhorp, G., and Pawson, R. (2011). Protocol – realist and meta-narrative evidence synthesis: Evolving standards (rameses). *BMC Med. Res. Method.* 11, 115. doi: 10.1186/1471-2288-11-115

Heintze, K., and Petersen, K. U. (2013). The case of drug causation of childhood asthma: antibiotics and paracetamol. *Eur. J. Clin. Pharmacol.* 69, 1197–1209. doi: 10.1007/s00228-012-1463-7

Henderson, A. J., and Shaheen, S. O. (2013). Acetaminophen and asthma. *Paediatric Respir. Rev.* 14, 9–16. doi: 10.1016/j.prrv.2012.04.004

Herxheimer, A. (2012). Pharmacovigilance on the turn? Adverse reactions methods in 2012. *Br. J. Gen. Pract.* 62, 400–401. doi: 10.3399/bjgp12X653453

Hill, A. B. (1965). The environment and disease: association or causation? *Proc. R. Soc. Med.* 58, 295–300.

Holman, B., and Bruner, J. P. (2015). The problem of intransigently biased agents. *Philosophy Sci.* 82, 956–968. doi: 10.1086/683344

Holman, B., and Geislar, S. (2019). Sex drugs and corporate ventriloquism: how to evaluate science policies intended to manage industry-funded bias. *Philosophy Sci.* doi: 10.1086/699713

Holman, B. H. (2015). *The fundamental antagonism: science and commerce in medical epistemology* (Ann Arbor: University of California, Irvine).

Horton, R. (2004). Vioxx, the implosion of Merck, and aftershocks at the FDA. *Lancet* 364, 1995–1996. doi: 10.1016/S0140-6736(04)17523-5

Howick, J. H. (2011). *The Philosophy of Evidence-Based Medicine* (Chichester: Blackwell).

Howson, C., and Urbach, P. (2006). *Scientific Reasoning* Vol. 3 (Chicago and La Salle: Open Court).

Ioannidis, J. P. (2016). Evidence-based medicine has been hijacked: a report to David Sackett. *J. Clin. Epidemiol.* 73, 82–86. doi: 10.1016/j.jclinepi.2016.02.012

Jüni, P., Nartey, L., Reichenbach, S., Sterchi, R., Dieppe, P. A., and Egger, M. (2004). Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. *Lancet* 364, 2021–2029. doi: 10.1016/S0140-6736(04)17514-4

Kahlert, J., Gribsholt, S. B., Gammelager, H., Dekkers, O. M., and Luta, G. (2017). Control of confounding in the analysis phase–an overview for clinicians. *Clin. Epidemiol.* 9, 195–204. doi: 10.2147/CLEP.S129886

Karch, F. E., and Lasagna, L. (1977). Toward the operational identification of adverse drug reactions. *Clin. Pharmacol. Ther.* 21, 247–254. doi: 10.1002/cpt1977213247

Karimi, M., Mirzaei, M., and Ahmadieh, M. H. (2006). Acetaminophen use and the symptoms of asthma, allergic rhinitis and eczema in children. *Iranian J. Allergy Asthma Immunol.* 5, 63–67.

Kastner, M., Tricco, A. C., Soobiah, C., Lillie, E., Perrier, L., Horsley, T., et al. (2012). What is the most appropriate knowledge synthesis method to conduct a review? Protocol for a scoping review. *BMC Med. Res. Method.* 12, 114. doi: 10.1186/1471-2288-12-114

Kastner, M., Antony, J., Soobiah, C., Straus, S. E., and Tricco, A. C. (2016). Conceptual recommendations for selecting the most appropriate knowledge synthesis method to answer research questions related to complex evidence. *J. Clin. Epidemiol.* 73, 43–49. doi: 10.1016/j.jclinepi.2015.11.022

Kelly, F. (1999). Gluthathione: in defence of the lung. *Food Chem. Toxicol.* 37, 963–966. doi: 10.1016/S0278-6915(99)00087-3

Knowledge Base workgroup of the Observational Health Data Sciences and Informatics (2017). Large-scale adverse effects related to treatment evidence standardization (laertes): an open scalable system for linking pharmacovigilance evidence sources with clinical data. *J. Biomed. Semant.* 8, 15. doi: 10.1186/s13326-017-0115-3

Kourounakis, A. P., Rekka, E. A., and Kourounakis, P. N. (1997). Antioxidant activity of guaiazulene and protection against paracetamol hepatotoxicity in rats. *J. Pharm. Pharmacol.* 49, 938–942. doi: 10.1111/j.2042-7158.1997.tb06140.x

Koutkias, V. G., and Jaulent, M. C. (2015). Computational approaches for pharmacovigilance signal detection: toward integrated and semantically-enriched frameworks. *Drug Saf.* 38, 219–232. doi: 10.1007/s40264-015-0278-8

Koutkias, V. G., Louët, A. L. L., and Jaulent, M. C. (2017). Exploiting heterogeneous publicly available data sources for drug safety surveillance: computational framework and case studies. *Expert Opin. Drug Saf.* 16, 113–124. doi: 10.1080/14740338.2017.1257604

Krauth, D., Woodruff, T. J., and Bero, L. (2013). Instruments for assessing risk of bias and other methodological criteria of published animal studies: a systematic review. *Environ. Health Perspect.* 121, 985–992. doi: 10.1289/ehp.1206389

Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K., et al. (2005). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am. J. Epidemiol.* 163, 262–270. doi: 10.1093/aje/kwj047

La Caze, A. (2009). Evidence-based medicine must be. *J. Med. Philosophy* 34, 509–527. doi: 10.1093/jmp/jhp034

Landes, J., and Williamson, J. (2016)."Objective Bayesian nets from consistent datasets." In *Proceedings of MaxEnt*, Eds. A. Giffin, and K. H. Knuthvol. vol. 1757 (Potsdam, NY, U.S.: AIP Publishing), 020007–1 – 020007–8. doi: 10.1063/1.4959048

Landes, J., Osimani, B., and Poellinger, R. (2018). Epistemology of Causal Inference in Pharmacology. *Eur. J. Philosophy Sci.* 8, 3–49. doi: 10.1007/s13194-017-0169-1

Landes, J. (2018). "An Evidence-Hierarchical Decision Aid for Ranking in Evidence-Based Medicine," in *Uncertainty in Pharmacology: Epistemology, Methods and Decisions*. Eds. B. Osimani, and A. La Caze (Boston Studies in Philosophy of Science: Springer), 31. doi: 10.1007/978-3-030-29179-2_11

LeBel, E. P., Vanpaemel, W., McCarthy, R., Earp, B., and Elson, M. (2018). A unified framework to quantify the trustworthiness of empirical research. *AMPPS.* 1 (3), 389–402. doi: 10.1177/2515245918787489

Lesko, S. M., and Mitchell, A. A. (1999). The safety of acetaminophen and ibuprofen among children younger than two years old. *Pediatrics* 104, e39. doi: 10.1542/peds.104.4.e39

Lesko, S. M., Louik, C., Vezina, R. M., and Mitchell, A. A. (2002). Asthma morbidity after the short-term use of ibuprofen in children. *Pediatrics* 109, e20. doi: 10.1542/peds.109.2.e20

Lindley, D. V. (2000). The Philosophy of Statistics. *J. R. Stat. Soc. Ser. D (The Statistician)* 49, 293–337. doi: 10.1111/1467-9884.00238

Lucas, P. J., Baird, J., Arai, L., Law, C., and Roberts, H. M. (2007). Worked examples of alternative methods for the synthesis of qualitative and quantitative research in systematic reviews. *BMC Med. Res. Method.* 7, 4. doi: 10.1186/1471-2288-7-4

Lundh, A., and Bero, L. (2017). The ties that bind. *Br. Med. J.* 356. doi: 10.1136/bmj.j176

Lundh, A., Lexchin, J., Mintzes, B., Schroll, J. B., and Bero, L. (2017). Industry sponsorship and research outcome. *Cochrane Lib.* doi: 10.1002/14651858.MR000033.pub3

Martinez-Gimeno, A., and García-Marcos, L. (2013). The association between acetaminophen and asthma: should its pediatric use be banned? *Expert Rev. Respir. Med.* 7, 113–122. doi: 10.1586/ers.13.8

McCarron, C. E., Pullenayegum, E. M., Thabane, L., Goeree, R., and Tarride, J. E. (2010). The importance of adjusting for potential confounders in bayesian hierarchical models synthesising evidence from randomised and non-randomised studies: an application comparing treatments for abdominal aortic aneurysms. *BMC Med. Res. Method.* 10, 64. doi: 10.1186/1471-2288-10-64

McKeever, T. M., Lewis, S. A., Smit, H. A., Burney, P., Britton, J. R., and Cassano, P. A. (2005). The Association of Acetaminophen, Aspirin, and Ibuprofen with Respiratory Disease and Lung Function. *Am. J. Respir. Crit. Care Med.* 171, 966–971. doi: 10.1164/rccm.200409-1269OC

Micheli, L., Cerretani, D., Fiaschi, A., Giorgi, G., Romeo, M., and Runci, F. (1994). Effect of acetaminophen on glutathione levels in rat testis and lung. *Environ. Health Perspect.* 102, 63–64. doi: 10.1289/ehp.94102s963

Montgomery, A. A., Peters, T. J., and Little, P. (2003). Design, analysis and presentation of factorial randomised controlled trials. *BMC Med. Res. Method.* 3, 26. doi: 10.1186/1471-2288-3-26

Mumford, S., and Anjum, R. L. (2011). *Getting causes from powers* (Oxford: Oxford University Press).

Murtas, R., Dawid, A. P., and Musio, M. (2017). New bounds for the probability of causation in mediation analysis. https://arxiv.org/abs/1706.04857

Naranjo, C. A., Busto, U., Sellers, E. M., Sandor, P., Ruiz, I., Roberts, E., et al. (1981). A method for estimating the probability of adverse drug reactions. *Clin. Pharmacol. Ther.* 30, 239–245. doi: 10.1038/clpt.1981.154

Nassini, R., Materazzi, S., AndrÃ¨, E., Sartiani, L., Aldini, G., Trevisani, M., et al. (2010). Acetaminophen, *via* its reactive metabolite N-acetyl-p-benzo-quinoneimine and transient receptor potential ankyrin-1 stimulation, causes neurogenic inflammation in the airways and other tissues in rodents. *FASEB J.* 24, 4904–4916. doi: 10.1096/fj.10-162438

Neapolitan, R. E. (2003). *Learning Bayesian Networks* (Upper Saddle River: Pearson).

Newson, R., Shaheen, S., Chinn, S., and Burney, P. (2000). Paracetamol sales and atopic disease in children and adults: an ecological analysis. *Eur. Respir. J.* 16, 817–823. doi: 10.1183/09031936.00.16581700

Onakpoya, I. J., Heneghan, C. J., and Aronson, J. K. (2016). Worldwide withdrawal of medicinal products because of adverse drug reactions: a systematic review and analysis. *Crit. Rev. In Toxicol.* 46, 477–489. doi: 10.3109/10408444.2016.1149452

Osimani, B. (2014). Hunting side effects and explaining them: should we reverse evidence hierarchies upside down? *Topoi* 33, 295–312. doi: 10.1007/s11245-013-9194-7

Osimani, B. (2019). "Social games and epistemic losses: reliability and higher order evidence in medicine and pharmacology," in *Uncertainty in Pharmacology: Epistemology, Methods, and Decisions*. Eds. A. La Caze, and B. Osimani ((Springer), Boston Studies in Philosophy of Science).

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. 1 (Cambridge: Cambridge University Press).

Poellinger, R. (2019). "Analogy-Based Inference Patterns in Pharmacological Research," in *Uncertainty in Pharmacology: Epistemology, Methods, and Decisions*. Eds. A. La Caze, and B. Osimani ((Springer), Boston Studies in Philosophy of Science).

Preston, T. A. (1988). "DES and the elusive goal of drug safety," in *Worse than the Disease: Pitfalls of Medical Progress*. Ed. D. B. Dutton (Cambridge: Cambridge University Press), 31–90. chap. 3.

Price, K. L., Amy Xia, H., Lakshminarayanan, M., Madigan, D., Manner, D., Scott, J., et al. (2014). Bayesian methods for design and analysis of safety trials. *Pharm. Stat.* 13, 13–24. doi: 10.1002/pst.1586

Puhan, M. A., Singh, S., Weiss, C. O., Varadhan, R., and Boyd, C. M. (2012). A framework for organizing and selecting quantitative approaches for benefit-harm assessment. *BMC Med. Res. Method.* 12, 173. doi: 10.1186/1471-2288-12-173

Reichenbach, H. (1956). *The Direction of Time* (Berkeley and Los Angeles: University of California Press).

Rising, K., Bacchetti, P., and Bero, L. (2008). Reporting bias in drug trials submitted to the food and drug administration: review of publication and presentation. *PloS Med.* 5, e217. doi: 10.1371/journal.pmed.0050217

Romero, F. (2016). Can the behavioral sciences self-correct? A social epistemic study. *Stud. Hist. Philosophy Sci. Part A* 60, 55–69. doi: 10.1016/j.shpsa.2016.10.002

Ryan, P., Suchard, M. A., Schuemie, M., and Madigan, D. (2013). Learning from epidemiology: interpreting observational database studies for the effects of medical products. *Stat. Biopharm. Res.* 5, 170–179. doi: 10.1080/19466315.2013.791638

Sanderson, S., Tatt, I. D., and Higgins, J. P. (2007). Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *Int. J. Epidemiol.* 36, 666–676. doi: 10.1093/ije/dym018

Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H., and Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* 20, 512–522. doi: 10.1097/EDE.0b013e3181a663cc

Shaheen, S. O., Sterne, J. A. C., Songhurst, C. E., and Burney, P. G. J. (2000). Frequent paracetamol use and asthma in adults. *Thorax* 55, 266–270. doi: 10.1136/thorax.55.4.266

Shaheen, S. O., Newson, R. B., Sherriff, A., Henderson, A. J., Heron, J. E., Burney, P. G. J., et al. (2002). Paracetamol use in pregnancy and wheezing in early childhood. *Thorax* 57, 958–963. doi: 10.1136/thorax.57.11.958

Shaheen, S., Potts, J., Gnatiuc, L., Makowska, J., Kowalski, M. L., Joos, G., et al. (2008). The relation between paracetamol use and asthma: a GA2)LEN European case-control study. *Eur. Respir. J.* 32, 1231–1236. doi: 10.1183/09031936.00039208

Shao, K., and Shapiro, A. J. (2019). A web-based system for bayesian benchmark dose estimation. *Environ. Health Perspect.* 126, 017002. doi: 10.1289/EHP1289

Shinkins, B., Yang, Y., Abel, L., and Fanshawe, T. R. (2017). Evidence synthesis to inform model-based cost-effectiveness evaluations of diagnostic tests: a methodological review of health technology assessments. *BMC Med. Res. Method.* 17, 56. doi: 10.1186/s12874-017-0331-7

Shrier, I., and Platt, R. W. (2008). Reducing bias through directed acyclic graphs. *BMC Med. Res. Method.* 8, 70. doi: 10.1186/1471-2288-8-70

Smith, L. J., Anderson, J., Shamsuddin, M., and Hsueh, W. (1990). Effect of fasting on hyperoxic lung injury in mice: the role of glutathione. *Am. Rev. Respir. Dis.* 141, 141–149. doi: 10.1164/ajrccm/141.1.141

Song, F., Parekh-Bhurke, S., Hooper, L., Loke, Y. K., Ryder, J. J., Sutton, A. J., et al. (2009). Extent of publication bias in different categories of research cohorts: a meta-analysis of empirical studies. *BMC Med. Res. Method.* 9, 79. doi: 10.1186/1471-2288-9-79

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search* (Cambridge: MIT press).

Sprenger, J., and Stegenga, J. (2017). Three arguments for absolute outcome measures. *Philosophy Sci.* 84, 840–852. doi: 10.1086/693930

Sprenger, J. (2019). The objectivity of subjective bayesianism. *Eur. J. Philosophy Sci.* doi: 10.1007/s13194-018-0200-1

Stare, J., and Maucort-Boulch, D. (2016). Odds ratio, hazard ratio and relative risk. *Adv. In Method. Statistics/Metodoloski zvezki* 13, 59–67.

Stausberg, J., and Hasford, J. (2011). Drug-related admissions and hospital-acquired adverse drug events in germany: a longitudinal analysis from 2003 to 2007 of ICD-10-coded routine data. *BMC Health Serv. Res.* 11, 134. doi: 10.1186/1472-6963-11-134

Stegenga, J. (2014). "Herding QATs: Quality assessment tools for evidence in medicine," in *Classification, Disease and Evidence* (Dordrecht: Springer), 193–211. doi: 10.1007/978-94-017-8887-8_10

Sterne, J. A., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., et al. (2016). ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 355, 1–7. doi: 10.1136/bmj.i4919

Stewart, G. B., Higgins, J. P. T., Schünemann, H., and Meader, N. (2015). The use of bayesian networks to assess the quality of evidence from research synthesis: 1. *PloS One* 10, e0114497. doi: 10.1371/journal.pone.0114497

Suppes P. (Ed.) (1970). *A Probabilistic Theory of causality* (Amsterdam: North-Holland Pub. Co.).

Sutton, A. J., and Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Stat. Methods In Med. Res.* 10, 277–303. doi: 10.1177/096228020101000404

Sutton, A. J., Cooper, N. J., Abrams, K. R., Lambert, P. C., and Jones, D. R. (2005). A bayesian approach to evaluating net clinical benefit allowed for parameter uncertainty. *J. Clin. Epidemiol.* 58, 26–40. doi: 10.1016/j.jclinepi.2004.03.015

Swaen, G., and van Amelsvoort, L. (2009). A weight of evidence approach to causal inference. *J. Clin. Epidemiol.* 62, 270–277. doi: 10.1016/j.jclinepi.2008.06.013

Talbott, W. (2011). "Bayesian epistemology," in *Stanford Encyclopedia of Philosophy*. Ed. E. N. Zalta (Stanford: Metaphysics Research Lab, Stanford University). Summer 2011 edn.

Teira, D., and Reiss, J. (2013). Causality, impartiality and evidence-based policy. *Mech. Causality Biol. Econ.* (Dordrecht), 207–224.

Teira, D. (2013). On the impartiality of early British clinical trials. *Stud. History Philosophy Sci. Part C: Stud. History Philosophy Biol. Biomed. Sci.* 44, 412–418. doi: 10.1016/j.shpsc.2013.05.003

Thomas, J., and Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Med. Res. Method.* 8, 45. doi: 10.1186/1471-2288-8-45

Tricco, A. C., Soobiah, C., Antony, J., Cogo, E., MacDonald, H., Lillie, E., et al. (2016a). A scoping review identifies multiple emerging knowledge synthesis methods, but few studies operationalize the method. *J. Clin. Epidemiol.* 73, 19–28. doi: 10.1016/j.jclinepi.2015.08.030

Tricco, A. C., Antony, J., Soobiah, C., Kastner, M., MacDonald, H., Cogo, E., et al. (2016b). Knowledge synthesis methods for integrating qualitative and quantitative data: a scoping review reveals poor operationalization of the methodological steps. *J. Clin. Epidemiol.* 73, 29–35. doi: 10.1016/j.jclinepi.2015.12.011

US Department of Health and Human Services, Office of Disease Prevention and Health Promotion (2014). National action plan for adverse drug event prevention. Accessed 17. Oct 2017.

van den Berg, T., Heymans, M. W., Leone, S. S., Vergouw, D., Hayden, J. A., Verhagen, A. P., et al. (2013). Overview of data-synthesis in systematic reviews of studies on outcome prediction models. *BMC Med. Res. Method.* 13, 42. doi: 10.1186/1471-2288-13-42

Vandenbroucke, J. P., and Psaty, B. M. (2008). Benefits and risks of drug treatments: how to combine the best evidence on benefits with the best data about adverse effects. *Jama* 300, 2417–2419. doi: 10.1001/jama.2008.723

Varallo, F. R., Planeta, C. S., Herdeiro, M. T., and de Carvalho Mastroianni, P. (2017). Imputation of adverse drug reactions: causality assessment in hospitals. *PloS One* 12, e0171470. doi: 10.1371/journal.pone.0171470

Varner, A. E., Busse, W. W., and Lemanske, R. F. (1998). Hypothesis: decreased use of pediatric aspirin has contributed to the increasing prevalence of childhood asthma. *Ann. Allergy Asthma Immunol.* 81, 347–351. doi: 10.1016/S1081-1206(10)63127-4

Warren, F. C., Abrams, K. R., Golder, S., and Sutton, A. J. (2012). Systematic review of methods used in meta-analyses where a primary outcome is an adverse or unintended event. *BMC Med. Res. Method.* 12, 64. doi: 10.1186/1471-2288-12-64

Watson, S., Chandler, R. E., Taavola, H., Härmark, L., Grundmark, B., Zekarias, A., et al. (2018). Safety concerns reported by patients identified in a collaborative signal detection workshop using vigibase: results and reflections from lareb and uppsala monitoring centre. *Drug Saf.* 41, 203–212. doi: 10.1007/s40264-017-0594-2

Wells, G. A., Shea, B., O'Connell, D., Peterson, J., Welch, V., Losos, M., et al. (2018). The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses.

Wood, L., Egger, M., Gluud, L. L., Schulz, K. F., Jüni, P., Altman, D. G., et al. (2008). Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ* 336, 601–605. doi: 10.1136/bmj.39465.451748.AD

Woodward, J. (2003). "Oxford Studies in the Philosophy of Science," in *Making Things Happen: A Theory of Causal Explanation* (Oxford: Oxford University Press).

Wu, T. Y., Jen, M. H., Bottle, A., Molokhia, M., Aylin, P., Bell, D., et al. (2010). Ten-year trends in hospital admissions for adverse drug reactions in England 1999-2009. *J. R. Soc. Med.* 103, 239–250. doi: 10.1258/jrsm.2010.100113

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.