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

Francesco De Pretis¹,²†, Jürgen Landes³† and Barbara Osimani¹,³*†

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-Universität München, München, Germany

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; Staab 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 (Herzheime, 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 (Elhammer 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 (Swan 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 probabilistic (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

¹Note 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).
The hypothesis of interest is that “Drug $D$ 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 $\varnothing$ 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 $\varnothing$ is a root node. The causal indicator are parents of report nodes which mediate causal inference from the concrete data (the reports) towards the causal hypothesis.

![Figure 1](image_url) | 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 REPs). In this paper, we focus exclusively on the REL modulator which we split into a number of concepts, see Evidential Modulators: Study Design.

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.

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

---

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

3They are extensively addressed in (Poellinger, 2019) and will be the focus of a rejoinder to this paper.
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 $D$ 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) $D$ causes $E$, 2) $D$ hinders $E$, and 3) $D$ does not cause $E$, we divide it here for simplicity’s sake into two alternative hypotheses: 1) $D$ causes $E$ and 2) $D$ 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 $D©E$, or simply ©.5

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, $P(\text{Ind} \mid \circledast) > P(\text{Ind}) > P(\text{Ind} \mid \bigtriangledown)$ but they are not entailed by it.

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

---

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

5We 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.

6Bradford 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 \( D \) and \( E \) stand in a difference-making relationship, then changes in \( D \) 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 \( D \) 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 \( D \) causes \( E \), or in cases where \( E \) causes \( D \), or when \( D \) and \( E \) have a common cause. All that matters is whether there is a (certain) systematic relationship between \( D \) and \( E \). RoG, DR, PD are semantically and epistemically related and we refer to them as statistical “black-box” indicators, denoted by \( \Sigma \).

**Mechanistic Knowledge (M)**

\( M \) represents the proposition: “there is a mechanism for \( D \) 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

---

\(^7\)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 D and E stand in the right temporal relationship (time course), which can refer to temporal order, distance, or duration. If D causes E, T must hold (as a necessary condition): © ⇒ T. T remains an imperfect indicator, nevertheless, because temporal precedence is also compatible with ¬(D causing E) when 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 indicators9 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 causation typical of the causal graph methodology developed by Pearl (2000) and Glymour and colleagues (Spirtes et al., 2000) see also Woodward (2003).10

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/...
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,P)I**

The main instruments to isolate the putative causal link \(D \oplus E\) 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 \(\Delta\) (and not even about any of the \(\Sigma\) 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 \(A\) produces an improvement of symptoms, 2) or it is drug \(B\) that worsens the situation by the same amount, 3) or else, both drug \(A\) and \(B\) 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 \(\Delta\) 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).

**Mechanic 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

---

14Our 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 to calculate 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.

15Other 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.

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

17In 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.

18In 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.  

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

\[
P(M \mid ©) = P(T \mid ©) = P(\bar{\Delta} \mid \bar{©}) = 1
\]

\[
P(PD \mid DR) = P(DR \mid RoG) = P(T \mid M) = 1,
\]

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

Similarly, the probability of there being probabilistic dependence between \( D \) 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 \( \Sigma \) 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, \( P(PD \mid ©DR) \) is equal to:

\[
P(\text{curves exhibiting } PD, © \text{and } DR) \]

\[
P(\text{curves exhibiting } © \text{and } DR) \]

Difference Making

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

\[
P(\Delta \mid ©) = 1 \quad \text{and} \quad P(\Delta \mid \bar{©}) = 0.
\]

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 \( D \) to \( E \), whether causal or not, or by there being a causal connection between \( D \) and \( E \). This is because, if there exists a mechanism from \( D \) to \( E \), then \( D \) must be prior to \( E \).\(^{22}\)

Also if \( D \) causes \( E \), then \( D \) must be prior to \( E \) as well. So:

\[
P(T \mid M, ©) = P(T \mid M, \bar{©}) = P(T \mid \bar{M}, ©) = 1.
\]

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 \( D \). We are hence indifferent over whether there is time precedence or not. So,

\[
P(T \mid \bar{M}, \bar{©}) = 0.5.
\]

Mechanisms

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

\(^{21}\)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 to be 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).

\(^{22}\)The fact that when \( D \) causes \( E \), in an individual or at a population level, it must come before \( E \), does not obviously entail that the drug \( D \) cannot also be taken after the event. We thank an anonymous reviewer for pointing this out to us.
from \( D \) to \( E \), even if \( D \) 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 \( \Sigma \) 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 \( \text{Rep} \) (for report). These report variables as well as the modulator variables (see Statistical Evidence for the \( \Sigma \)-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 \( \Sigma \)-Indicators**

We assume that every observational study yields information about one \( \Sigma \) indicator only: i.e., each \( \text{Rep} \) node only has one \( \Sigma \) 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 \( \Sigma \) indicator only (\( \text{PD} \)).

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 \( \Sigma \)-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.

**Evidence of Difference-Making**

RCTs inform us about the difference-making indicator of causation and whether there is time precedence. For each study,
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 to point 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, for every mechanism for which there is evidence. The Mi come as hypotheses about mechanisms between D and E. Each mechanism Mi may be broken down into further bits of the mechanism; denoted here by μi,k. Concrete data about these bits is denoted by Repmi,k, see Figure 7 for an illustration.

As for the reports feeding into the Σ set or to Δ, also the Repmi,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.

**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) 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

---

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

---

**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,M,…M, their constitutive sub-mechanisms (μ) and concrete evidence (Rep). Dotted edges are present, if and only if two M share parts of their mechanisms. Sub-mechanisms nodes (μ) 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.
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 $P(\otimes)$: 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 $\Delta$ 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 $\bar{x}$ to denote the values of the pertinent modulators here and in the following formulae.$^{25}$

$$P(ES = 0 | \bar{x}, \Delta) = 1 - P(ES = 1 | \bar{x}, \Delta) = 1 - (1 - \frac{SB}{10}) \cdot (1 - 0.5 \cdot \| \text{avg}(A, SS, D, B, R, Pl) \|)$$

$$= 1 - \left( 0.9 \cdot \left( 0.5 + \frac{0.5 + 1 + 0 + 1 + 0 + 0.5}{12} \right) \right) = \frac{13}{40}$$

$$P(ES = 0 | \bar{x}, \Delta) = 1 - P(ES = 1 | \bar{x}, \Delta) = 1 - (1 - \frac{SB}{10}) \cdot (0.5 \cdot \| \text{avg}(A, SS, B, R, Pl) \|)$$

$$= 1 - 0.9 \cdot \left( 0.5 \cdot \left( 1 - \frac{0.5 + 1 + 1 + 1 + 0.5}{5} \right) \right) = \frac{91}{100}.$$  

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,
Mechanistic Evidence

To focus the exposition, we only consider two possible mechanisms (M₁ and M₂) by which paracetamol may cause asthma.

M₁: Paracetamol is metabolised to NAPQI (N-acetyl-p-benzoquinone imine) (μ₁,3), NAPQI stimulates transient receptor potential ankyrin-1 (TRPA1) (μ₁,2) [reported in Nassini et al. (2010)] and TRPA1 causes airway neurogenic inflammation (μ₁,1) [reported in Nassini et al. (2010)].

M₂: Paracetamol depletes Gluthathione (μ₂,3) [reported in Micheli et al. (1994); Kourounakis et al. (1997)], low levels of Gluthathione cause oxidative stress hyperresponsiveness in the airways (μ₂,2) [reported in Smith et al. (1990); Kelly (1999)].

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

\[ P(M_1 \mid M) = 0.7 \quad P(M_1 \mid \overline{M}) = 0 \]
\[ P(M_2 \mid M) = 0.8 \quad P(M_2 \mid \overline{M}) = 0. \]

We assessed M₁ and M₂ to be likely, if M holds; M₁ was assessed to be the more likely of the two. If M does not hold, then all Mᵢ have to fail to hold and are hence assigned zero probability.

We now turn to setting conditional probabilities of the μᵢ, given Mᵢ and given M. First, recall that M entails μᵢ, and hence

\[ P(\mu_{i,1} \mid M_1) = 1 \]
\[ P(\mu_{i,2} \mid M_1, \mu_{i,1}) = 1 \]
\[ P(\mu_{i,3} \mid M_1, \mu_{i,1}, \mu_{i,2}) = 1. \]

\[ \mu_{i,1}, \mu_{i,2}, \mu_{i,3} \] and \( \overline{M}_i \) are, when taken together, logically inconsistent. So,

\[ P(\mu_{i,1} \mid \overline{M}_i, \mu_{i,1}, \mu_{i,2}) = 0. \]

If \( \overline{M}_i \) fails to hold, then we are indifferent about \( \mu_{i,1} \) and \( \mu_{i,2} \) independently of \( \mu_{i,3} \) (respectively \( \mu_{i,3} \)).

\[ P(\mu_{i,2} \mid \overline{M}_i, \mu_{i,1}) = P(\mu_{i,2} \mid \overline{M}_i, \mu_{i,1}, \mu_{i,2}) = 0.5 \]
\[ P(\mu_{i,3} \mid \overline{M}_i, \mu_{i,1}, \mu_{i,2}) = P(\mu_{i,3} \mid \overline{M}_i, \mu_{i,1}, \mu_{i,2}) = 0.5. \]

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

\[ P(\mu_{i,1} \mid \overline{M}_i) = 1. \]

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

\[ P(Rep_{1,2} \mu_{i,2} \mid \mu_{i,2}) = 0.91 \quad P(Rep_{1,2} \mu_{i,2} \mid \mu_{i,2}) = 0.09 \]
\[ P(Rep_{2,3} \mu_{i,3} \mid \mu_{i,3}) = 0.91 \quad P(Rep_{2,3} \mu_{i,3} \mid \mu_{i,3}) = 0.09 \]

We take the quotient \( P(Rep_{1,2} \mu_{i,2}) / P(Rep_{1,2} \mu_{i,2}) \) 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 \( \mu \) than under \( \overline{\mu} \). A Bayes factor of 9/1 ≈ 9.1 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 \( M_2 \):

\[ P(\mu_{i,1} \mid M_2) = 1 \quad P(\mu_{i,1} \mid \overline{M}_2) = 0.01 \]
\[ P(\mu_{i,2} \mid M_2, \mu_{i,2}) = 1 \quad P(\mu_{i,2} \mid M_2, \mu_{i,2}) = 0.5. \]

\[ P(\mu_{i,2} \mid M_2, \mu_{i,2}) \] is zero for the same reasons as \( P(\mu_{i,3} \mid M_1, \mu_{i,1}, \mu_{i,2}) \) 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

\[ P(Rep_{1,2} \mu_{i,1} \mid \mu_{i,1}) = 0.91 \quad P(Rep_{1,2} \mu_{i,1} \mid \mu_{i,1}) = 0.09 \]
\[ P(Rep_{2,2} \mu_{i,2} \mid \mu_{i,2}) = 0.91 \quad P(Rep_{2,2} \mu_{i,2} \mid \mu_{i,2}) = 0.09 \]
\[ P(Rep_{1,2} \mu_{i,2} \mid \mu_{i,2}) = 0.75 \quad P(Rep_{1,2} \mu_{i,2} \mid \mu_{i,2}) = 0.25 \]
\[ P(Rep_{2,2} \mu_{i,2} \mid \mu_{i,2}) = 0.91 \quad P(Rep_{2,2} \mu_{i,2} \mid \mu_{i,2}) = 0.09. \]
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:

\[
P(DR | E_{DR})
\]

e.g., 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 \(P(DR | E_{DR}, x)\). 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.,

### Table 2

| Evidence Type | Prior Probability of © | Posterior Probability of © |
|---------------|------------------------|---------------------------|
| No Evidence   | 0.0100                 | 0.0050                    |
| Mechanical Evidence | 0.0175                 | 0.0088                    |
| Micheli et al. (1994) | 0.0193                 | 0.0097                    |
| Kourounakis et al. (1997) | 0.0195                 | 0.0098                    |
| Kelly (1999) | 0.0196                 | 0.0098                    |
| Nassini et al. (2010a) | 0.0196                 | 0.0099                    |
| Nassini et al. (2010b) | 0.0198                 | 0.0099                    |
| Statistical Evidence discussed above | 0.0072                 | 0.0036                    |
| Shaheen et al. (2002) | 0.1534                 | 0.0827                    |
| Statistical Evidence discussed in the Supplementary Material | 0.2238                 | 0.1254                    |
| Newson et al. (2000) | 0.0997                 | 0.0522                    |
| Lesko et al. (2002) | 0.3866                 | 0.2250                    |
| Barr et al. (2004) | 0.6397                 | 0.4690                    |
| McKeever et al. (2005) | 0.6445                 | 0.4742                    |
| Karimi et al. (2006) | 0.6446                 | 0.4743                    |
| Shaheen et al. (2008) | 0.6446                 | 0.4743                    |
| Amberbir et al. (2011) | 0.7055                 | 0.5437                    |
| Beasley et al. (2011) | 0.7160                 | 0.5564                    |

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

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
