Euro Jnl Phil Sci (2018) 8:3–49
DOI 10.1007/s13194-017-0169-1

ORIGINAL PAPER IN PHILOSOPHY OF SCIENCE

Epistemology of causal inference in pharmacology

Towards a framework for the assessment of harms

Jürgen Landes[1] · **Barbara Osimani**[1] · **Roland Poellinger**[1]

Received: 15 February 2016 / Accepted: 17 January 2017 / Published online: 16 March 2017
© Springer Science+Business Media Dordrecht 2017

Abstract Philosophical discussions on causal inference in medicine are stuck in dyadic camps, each defending one kind of evidence or method rather than another as best support for causal hypotheses. Whereas Evidence Based Medicine advocates the use of Randomised Controlled Trials and systematic reviews of RCTs as gold standard, philosophers of science emphasise the importance of mechanisms and their distinctive informational contribution to causal inference and assessment. Some have suggested the adoption of a pluralistic approach to causal inference, and an inductive rather than hypothetico-deductive inferential paradigm. However, these proposals deliver no clear guidelines about how such plurality of evidence sources should jointly justify hypotheses of causal associations. We here develop such guidelines by first giving a philosophical analysis of the underpinnings of Hill's (1965) viewpoints on causality. We then put forward an evidence-amalgamation framework adopting a Bayesian net approach to model causal inference in pharmacology for the assessment of harms. Our framework accommodates a number of intuitions already expressed in the literature concerning the EBM vs. pluralist debate on causal inference, evidence hierarchies, causal holism, relevance (external validity), and reliability.

This work is supported by the European Research Council (grant 639276) and the Munich Center for Mathematical Philosophy (MCMP).

✉ Barbara Osimani
b.osimani@lmu.de

Jürgen Landes
juergen.landes@lrz.uni-muenchen.de

Roland Poellinger
r.poellinger@lmu.de

[1] Munich Center for Mathematical Philosophy, LMU Munich, Munich, Germany

Keywords Causation · Evidence · Bayesian epistemology · Scientific inference · Safety assessment in pharmacology · Risk · Bradford Hill criteria

1 Introduction: causal inference in pharmacology

Pharmacology blends science and technology in a very peculiar fashion. It works across levels of reality by directly intervening at the biochemical level only: whereas the direct domain of action of drug molecules is limited to protein receptors, the desired end-effects are clinically observable results. However, because the proteins with which the drug molecules interact are embedded in various, possibly interacting, biological pathways (metabolic, genetic, signal transduction), most end-effects are unpredictable.

Knowledge of these various interactions and the *biological laws* governing them, as well of the contingent *initial conditions* holding in any specific context is far from exhaustive and does not allow reliable prediction or causal inference. Hence, until recently, drug approval has mainly relied on a black-box methodology, grounded on hypothesis rejection. This paradigm, which has been mainly developed with the aim to minimise false positives in efficacy assessment (see Howick 2011; Teira 2011, for an historical-philosophical overview), puts several constraints on the kind of evidence which is allowed to inform causal inference, and severely hampers the integration of heterogenous, possibly inconclusive pieces of evidence. Indeed, by giving a strong precedence to specific methods for inferring causes (such as Randomised Controlled Trials), the standard approach implicitly emphasises some indicators of causality, such as “difference making” in contrast to others, and has difficulties in incorporating evidence of different sources (e.g. spontaneous reports, case series, comparative studies of various kinds) and levels (e.g., molecular data, clinical evidence, epidemiological studies) which cannot be accommodated under this heading. Whereas this paradigm is reasonable for the purpose of avoiding fraud, by eliminating as much as possible any source of confounding and bias, it is not adequate for the purpose of minimising harms of health interventions (see Osimani and Mignini 2015).

The methodological landscape is rapidly changing though: Bayesian methods are gaining ground in statistical analysis of trials, because of their ability to optimise the use of available evidence by incorporating historical (heterogenous) knowledge in the prior, allowing diverse types of evidence to be integrated in the probability function, and by providing a probabilistic measure of the hypothesis under investigation, hence allowing decisions under uncertainty. Such statistical techniques are gaining ground especially in safety trials and trials for fast track or so called “orphan” drugs.

Systems pharmacology takes a “holistic” approach to study the effects of drugs in the organism by focusing on interrelations, rather than the components, of the mechanisms leading to intended and unintended outcomes. Computational modeling of perturbated cellular mechanisms for instance, aims to provide insights into the variability and complexity of pharmacological effects in the organ system (see for instance Britton et al. 2013). Also, data-intensive and knowledge discovery techniques put all available and drug-outcome relevant data together, in order to possibly predict the end-effect of a given drug, by reconstructing the possible routes of action from the molecular to the phenotypic level (see Abernethy and Bai 2013; Tilman and Eberhardt 2014; Xie et al. 2009). There is still considerable uncertainty as to the inferential and predictive value of these approaches, especially when they are taken on their own. However, their specific epistemic contribution could be made more valuable in combination with other sources of knowledge.

The main focus of the present paper is in fact to provide a framework for the amalgamation of diverse kinds of evidence for causal inference, which is formally and epistemologically grounded. For this, we will adopt Bovens' and Hartmann's (Bovens and Hartmann 2003) proposal to use Bayesian confirmation theory in order to account for (and mathematically explain) some phenomena related to scientific inference; such as the confirmatory power of the coherence of the body of evidence, the epistemic interaction of consistency of measurements and reliability of information sources, as well as the modular contribution of different "lines of evidence" related to diverse observable consequences of the investigated hypothesis. We will adapt this framework to causal inference and consequently specify a concrete structure for that purpose. We will then illustrate its epistemic and heuristic virtues as an instrument for evidence amalgamation, in the context of causal inference of drug-induced harm.

Section 2 explains in more detail why standards for efficacy and safety assessments should not be the same. Section 3 focuses on the role of causal inference in such assessments and elaborates on Bradford Hill viewpoints on causality in order to provide a list of philosophically grounded causal indicators. Section 4 presents Bayesian epistemology as a valid alternative to current standards of evidence; in particular it relies on Bovens' and Hartmann's mathematical representation of scientific inference and adapts it to the specific problem of inferring causality in pharmacology. Section 5 presents the details of our model and illustrates how it functions by presenting two model calculations. Section 6 spells out the main virtues of our inferential model with respect to the standard view and other proposals for synthesising evidence; furthermore it sets the stage for further developments and practical implementation.

2 Efficacy and safety assessments in pharmacology

2.1 Why standards ought not be the same

The European Parliament and the European Council have recently changed the regulation of pharmacovigilance practice (Directive 2010/84/EU; regulation (EU) No 1235/2010, entered into force in July 2012), putting a special emphasis on joint efforts for what can be considered an information-based (rather than power-based) approach to pharmaceutical risk assessment. The related guidelines encourage the integration of information coming from different sources of safety signals (spontaneous case reports, literature, data mining, pharmaco-epidemiological studies, post-marketing trials, drug utilisation studies, non-clinical studies, late-breaking information; see also Herxheimer 2012). Yet, the methodological bases for implementing such policy are still shaky in that the standards for causal assessment of adverse drug reactions (ADRs) is still parasitic on the (statistical) methods developed to test drug efficacy (see also Senn ; Price et al. 2014; Osimani and Mignini 2015). In Osimani and Mignini (2015), a series of reasons have been provided for adopting asymmetric standards for safety and efficacy assessments. These mainly deal with the following facts: 1) because of pragmatic as well as epistemic reasons, in the case of risk assessment there is higher concern for false negatives (failing to detect possible causes of observed effects), rather than for false positives (failing to distinguish between spurious and genuine causes). This is mainly due to the fact that the drug is developed and tested with reference to the intended effect, whereas the detection of side-effects – apart from most common ones, which can be observed also in medium-sized samples – is mainly left to the post-marketing phase; 2) also, evidence accumulating in time may strongly point to the hypothesis of causal association between drug and observed harm, without nevertheless being conclusive. Hence, instruments of probabilistic causal assessment are needed which do not demand a clear-cut rejection or acceptance of investigated hypotheses; 3) heterogeneous pieces of evidence might jointly support a given hypothesis without being individually able to allow any significant inference; hence instruments for evidence amalgamation are needed for this purpose.

2.2 The decision problem

Drug licensing bodies, such as the Food and Drug Agency in the USA or the EMA in Europe, as well as national agencies such as the Bundesinstitut für Arzneimittel und Medizinprodukte in Germany, or the Medicines and Healthcare products Regulatory Agency in the UK, regularly face the problem of whether to approve a drug for treatment or not and the problem of whether or not to let a drug further circulate in the market when its safety profile is updated through the discovery of additional risks. Indeed any given drug is always approved “with reservation” (Osimani 2007).[1] The actions taken by the drug licensing body may have wide influence on public health and public finance as well as economic success of the drug’s manufacturer and its competitors. Intuitively, the normatively right action to take is to leave the drug in the market, provided that – on the basis of the available evidence – the expected utility of not withdrawing it exceeds the utility of withdrawing it.

The precautionary principle has been introduced in the pharmaceutical domain in order to account for the uncertainty arising in cases where suspicion arises about a new harm, possibly associated with the drug, but evidence cannot conclusively point to a causal connection between them. In fact, before its introduction in the legal

[1] In reality, the decision problem is not as black and white as presented here. There are further actions available to a drug licensing agency such as: restricting access to the drug to a subset of patients and adding further information to the package insert, such as black-box warnings. For ease of exposition, we shall here disregard further possible actions and only consider the black and white decision problem on whether or not to approve a drug or leave it in the market after its safety profile has changed.

system (and through various international agreements related to environmental law, see Osimani 2013) no preventive measure was possible without a scientific proof of the causal connection between suspected source of damage and expected harm. This is because liability and safety regulations are grounded on a clear causal connection between the agent deemed responsible for the hazard and the hazard itself. The precautionary principle relaxes this requirement in view of the good at stakes (health and environment) and of the radical uncertainty related to the possible unintended outcomes of human interventions on nature. In such cases, a well-founded suspicion may suffice to take action (withdraw the drug or restrict its usage), and the principle of proportionality applies. This means that the probability associated with the hypothesis of causal association may be as low as the expected harm is high (with respect to the expected benefit, Osimani et al. 2011).

Hence, the decision, e.g., *withdraw the drug or not*, will depend on some threshold which reflects the nature of the medication, the pharmaceutical environment (i.e., the availability of alternative treatments for the same condition), policy and ethical dimensions, as well as the perceived acceptability of the risk.

In order to avoid commitment (for the moment) to any of the many notions of causation offered by the philosophical-methodological literature[2] we use the formula: $D\copyright H$ in order to express the proposition: "D causes H", sometimes abbreviated by © when no ambiguities arise. Hence, by adopting the classical cost-effectiveness analysis formula, we can infer the probability threshold for causality p^*, at which the expected utility of withdrawing equals the expected utility of keeping the drug in the market.

Let w stand for the act of withdrawing the drug D from the market while $\neg w$ stands for not withdrawing D. The utility of (not) withdrawing given that © or $\neg$© holds is denoted by the two-place utility function U. At the probability threshold p^*, the expected utility for withdrawing the drug equals the expected utility for not withdrawing it. p^* can hence be obtained by solving

$$p^* \cdot U(w, \copyright) + (1 - p^*) \cdot U(w, \neg\copyright) = p^* \cdot U(\neg w, \copyright) + (1 - p^*) \cdot U(\neg w, \neg\copyright),$$

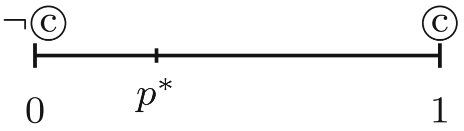
for p^*. Therefore, we can find p^* to be determined by the utilities

$$p^* = \frac{U(\neg w, \neg\copyright) - U(w, \neg\copyright)}{U(w, \copyright) - U(\neg w, \copyright) + U(\neg w, \neg\copyright) - U(w, \neg\copyright)}.$$

If the degree of belief in © is strictly greater than p^*, then the normatively correct decision is to withdraw the drug from the market. If the degree of belief is strictly less than p^*, then the normatively correct decision is to keep the drug in the market.

[2] We deliberately use the circled © as our relational predicate symbol here to not invoke any specific technical, theory-informed reading of this claim, yet. We precisely do think, though, that such an explication is in order – even more so since it is often neglected and rarely undertaken in the methodological literature.

Fig. 1 p^* partitions degrees of belief in © into two intervals



Hence, p^* partitions the continuum of degrees of belief between the two alternative hypotheses (© and ¬©) into two intervals, see Fig. 1.

There is a fact of the matter: either © is true or the opposite holds. So, in order to make the best decision it is necessary and sufficient to adopt degrees of belief which fall in the interval between p^* and the truth value of © – where a truth value of 1 stands for 'true' and a truth value of 0 stands for 'false'. Therefore, p^* allows for a certain margin of error; however, the chances to fall into the right interval get higher and higher the more evidence one takes into account: the more one "samples" from reality, the closer one's beliefs get to the truth (Edwards et al. 1963).[3]

A straightforward consequence of this state of affairs is that there is a need of instruments which allow a probabilistic assessment of the suspected causal link between drug and side-effect, by taking into account all available evidence at the time of decision. In particular, four desiderata are essential for a framework of causal assessment of drug induced harm:

1. It must allow for probabilistic hypothesis confirmation.
2. It must be able to incorporate heterogeneous kinds of data.
3. It must be able to integrate diverse types of inferential patterns, in order to optmise the epistemic import of available evidence.
4. The framework should be particularly focused on causal assessment in pharmacology and therefore consider the specific issues which arise in this context.

In this paper, we focus exclusively on how to develop a framework of this kind. The problems of i) how to determine the expected utilities of these harms and ii) how to determine the expected utility of benefits are outside the scope of this paper.

3 A framework for the assessment of harms in pharmacology, part 1: relata

In the following, we present our approach to causal inference for drug harms based on a formal Bayesian model of scientific inference. Our approach 1) identifies possible indicators of causality (observable consequences of the causal hypothesis) on the basis of the methodological and philosophical literature on causality, evidence, and causal inference; 2) embeds them in a topological framework of probabilistic

[3] This is the standard justification of inductive inference (Carnap 1947; Howson and Urbach 2006); we do not enter into the related philosophical debate here. However, nothing hinges on the particular philosophical position about truth.

dependencies and independencies grounded in assumptions regarding their reciprocal epistemic interconnections; 3) weakly orders some of these probabilistic dependencies as a function of their inferential strength with respect to the confirmation of causal hypotheses. Furthermore, the developed model is used to illustrate possible epistemic dynamics related to the interactions of its various components.

3.1 Causal indicators

In a much quoted paper devoted to causal assessment of environmental hazards (Hill 1965), the epidemiologist Sir Austin Bradford Hill identifies nine "viewpoints" which help detecting possible causes of observed risks. Hill does not consider these "viewpoints" (neither individually nor jointly) to provide sufficient and necessary conditions for causality, but they should "help us to make up our minds on the fundamental question – is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect?", see Hill (1965, p. 299). The nine "viewpoints" are the following: 1) strength of the association; 2) consistency; 3) specificity; 4) temporality; 5) biological gradient; 6) plausibility; 7) coherence; 8) experiment; 9) analogy.

1. *Strength of the association* refers to the observed relationship between a candidate cause and the putative risk: how much the former contributes to the latter, measured for instance by the ratio of relevant outcomes between exposed and unexposed group, or by regression coefficients. Particularly, Hill emphasises that although this kind of information may be causally opaque because of possible confounders, still causation may be conceded whenever such possible confounders cannot be reasonably identified (on grounds that one can use this as a heuristic basis for excluding them).
2. *Consistency* refers both to the convergence of the observed results in *different study settings and across different methods*, and to the stability of the association across *different background conditions* and circumstances. The former kind of consistency may be also referred to as methodological robustness, in that it is meant to provide a warrant that the observed results have not been produced by the studies themselves, and thereby exclude the possibility that they are a study artifact. The latter instead refers to the association holding in different contexts/ kinds of populations/background conditions, and is also known as "ontological robustness" (Wimsatt 1981, 2012).
3. *Specificity* refers to ideally one-to-one relationships between specific sources of hazards and specific kinds of harms. Hill makes for instance the example of different kinds of cancer sites (lung or nose vs. scrotum cancer) which specific populations (nickel refiners vs. chimney sweepers) are more likely to contract, depending on the kinds of chemicals they are exposed to.
4. *Temporality*. Excluding theoretical cases of backward causation, causes come before their effects. Hence, one would be inclined to infer causality when both a statistical association is present and the putative cause is observed to come prior to the effect. However, observational studies cannot guarantee perfect information on temporal order, since generally both the exposure to risk factors and the

development of disease extend through time; therefore attention should be paid to possible confounding factors and reverse causation.

5. *Biological gradient* refers to what is also called dose-response curve: an observed systematic relationship between the exposure and the strength of the observed effect: “For instance, the fact that the death rate from cancer of the lung rises linearly with the number of cigarettes smoked daily, adds a very great deal to the simpler evidence that cigarette smokers have a higher death rate than non-smokers”, see Hill (1965, p. 298).
6. *Biological plausibility* refers to the fact that knowledge of molecular mechanisms should also be considered, in addition to statistical knowledge, when assessing causal associations. On the other hand, causal association should not be dismissed on the grounds that the hypothesised mechanisms are implausible, since implausibility is relative to the state of the art, and this might be overturned by strong evidence to the contrary: “What is biologically plausible depends upon the biological knowledge of the day”, see Hill (1965, p. 298).
7. *Coherence* relates to how well the different pieces of evidence fit together. For instance, when evidence from animal studies and epidemiological studies point to the same hypothesis. Coherence also relates to how well the hypothesis fits with background knowledge: “the cause and effect interpretation of our data should not seriously conflict with the generally known facts of the natural history and biology of the disease”, see Hill (1965, p. 298). However, lack of coherence among the different kinds of studies, cannot totally nullify their evidential value.
8. *Experimental evidence*, if available, allows the scientist to isolate the factor under investigation from other possible confounders, and thereby provides the strongest support for causal hypotheses according to Hill and the epidemiological canon in general.
9. *Reasoning by analogy* is an additional source of potentially relevant information for the causal claim at hand. Hill only briefly mentions analogy as providing further support for the hypothesis under investigation.

Hill concedes that he is not providing a theoretical justification for his list of categories, but interestingly, his points of view on causality reflect many of the criteria discussed in the philosophical literature for several decades now. We shall go on to systematically gather philosophical support for Hill’s list in the next section.

3.2 Philosophical underpinnings of various indicators of causality

In the following, we discuss the rationales which epistemologically underpin Hill’s viewpoints on causality by appealing to the philosophical literature, and derive our list of indicators of causality for our formal framework. While there may be further indicators of causality (in pharmacology), we think that those presented below are the most pertinent ones discussed in the philosophical literature.[4] Figure 2 presents a

[4] Ioannidis (2016) examines the empirical evidence in support of Bradford Hill guidelines and deemphasise their importance. Indeed the influence of biases of various kinds may distort the genuine informative value of such indicators. In our framework, this aspect is explicitly taken into account by

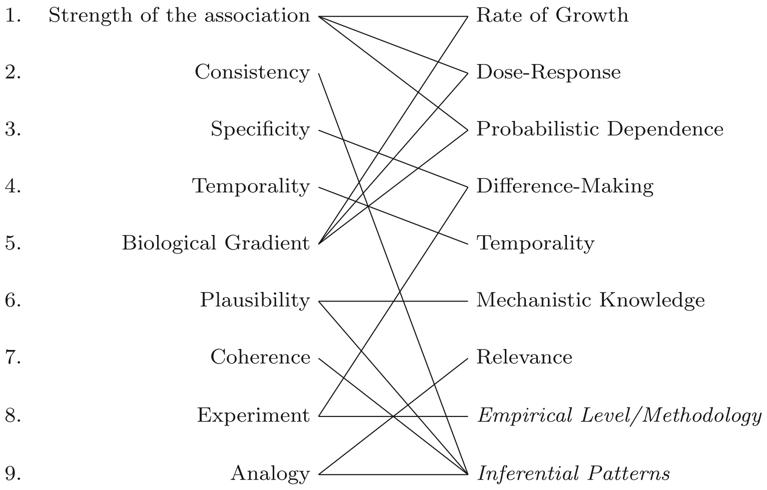


Fig. 2 Mapping Hill's nine viewpoints onto our framework, with indicators of causality and dimensions of scientific research (italicised) on the *right side*

mapping from Hill's list onto our framework which will be discussed in detail below. Not every viewpoint is mapped onto an *indicator of causality* in our sense – e.g., we locate *analogy* among the set of inferential patterns. The remainder of Section 3.2 is aligned with Hill's list, cf. Fig. 2.

3.2.1 Strength of association

Observed strength of association may mean different things. In biological systems, causes bring about putative effects, if specific combinations of background conditions hold and possibly hindering factors are absent. In contrast to causes which hold on a broad spectrum of background conditions, causes whose backgrounds conditions are rarely met – or whose hindering factors are common – will produce small effects in the treatment arm. Hence the effect-size reflects this kind of phenomenon, which philosophers have referred to as "stability" or "invariance", or "insensitivity" or "non-contingency" (Pearl 2000; Woodward 2003). This means that the effect size may reflect the relative non-contingency of the causal effect (independence from "supporting factors": Cartwright and Stegenga 2011; Stegenga 2015), rather than its causative force.

Woodward (2003, 2010) introduces stability as a feature of counterfactual dependence between putative cause and effect: the causal relationship remains intact under changes in the environment, and the severity of the changes "tests" the stability of the relationship – the more stable, the less sensitive to the specification of back-

providing separate lines of support for the confirmatory value of a given piece of evidence and its reliability.

ground parameters. Stability may indeed be considered as an indicator of causality in that it manifests the non-contingent relationship between the observed effect and the putative cause: in a certain sense it is an attenuated version of the necessity condition inherent to the pre-Humean conception of causality. Woodward as well as Pearl (Pearl 2000; Woodward 2003) consider stability as an essential quality of causal relations: the more unstable a relation is observed to be, the higher the likelihood that it is not genuinely causal at all in the end.

Strength of the association is also a function of how much the putative effect changes upon changes of the putative cause. This is measured for instance by the (regression) coefficient. In this respect, strength refers to the functional relationship itself, independently of its degree of universality. We will refer to a strong association in this sense as having a "high rate of growth".

Related to the rate of growth is also the "dose-response relationship". This states *whether* a systematic relation between changes in one variable and changes in the other holds in the first place. The existence of a dose-response model is an important indicator of causality especially because the detection of a clear dose-response curve (e.g., a logarithmic relation between treatment or exposure and observed effect values with small error terms) is best explained by positing a truly ontological influence structure along Reichenbach's principle: The variables are either directly causally related or correlated due to the presence of a common cause.[5]

Dose-response relationship and (high) rate of growth are epistemically related not only because the latter conceptually subsumes the former, but also because the higher the rate of growth, the more likely it is that both will be detected in observational or experimental studies. This is because, if the rate of growth is high, then a relationship between putative cause x and effect y (x and y are continuous variables) is more likely to be detected even in small samples/less favourable background conditions. In sum, the presence of a dose-response relationship $dy/dx \neq 0$ (for most x in the domain) suggests that there is a systematic relationship between the putative cause and the effect; the rate of growth says how strong this relationship is, e.g, how much dy/dx departs from 0. Figure 3 contrasts two exemplary rates of growth (high in graph 1a and low in graph 1b) with concrete dose-response curves (graphs 2a, 2b, and 2c). The dose-response relationship can be linear or – as illustrated in the example – nonlinear (in which case it can be monotonic or non-monotonic).[6]

[5] Drawing causal inferences from functional or statistical relations alone is a hard task and in many cases not feasible. If a functional description (like a structural equation) or a statistical connection (like a high measure of covariance) is available though (and has proven stable), it can be used for intervention and prediction – two hallmarks of causal knowledge. Although David Freedman criticises the Spirtes-Glymour-Scheines approach (Spirtes et al. 2000) towards automatically inferring causal claims from raw data, he points precisely to the practical use of formal dose-response relations when he writes that *"[t]hree possible uses for regression equations are (i) to summarise data, or (ii) to predict values of the dependent variable, or (iii) to predict the results of interventions"* (Freedman 1997, p. 62).

[6] This has implications for both causal inference as well as intervention and prediction: the more complex the functional relationship, the more difficult it is to detect it and to accurately represent it; and therefore the higher the risk of false prediction and inadequate intervention (Steel 2008). However, the framework presented here focuses on detecting causes rather than on using causal knowledge for prediction and intervention.

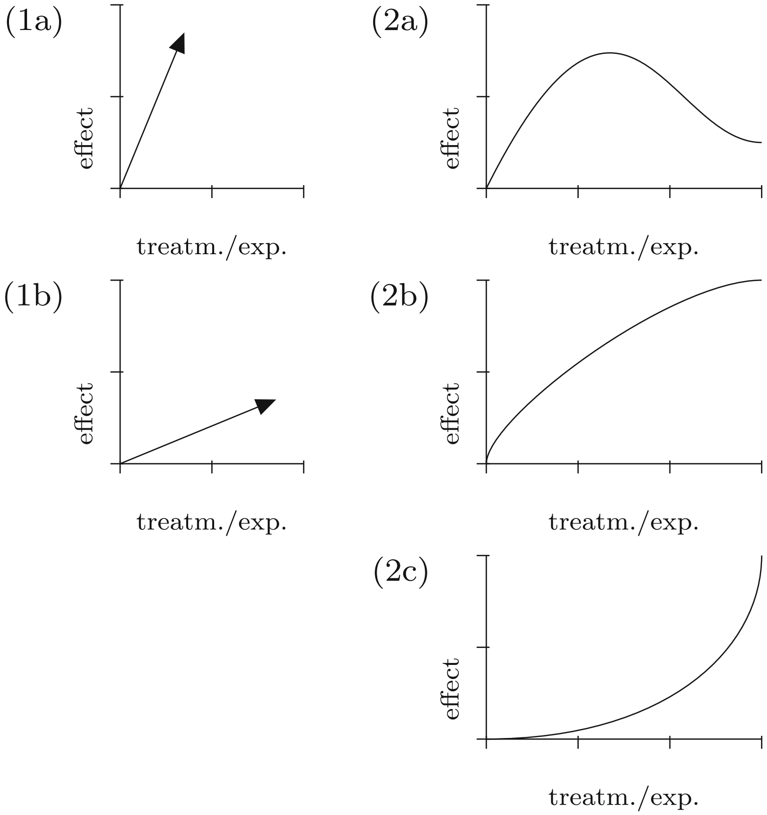


Fig. 3 Examples of different rates of growth (visualised as pointers) and different dose-response relationships (visualised as dose-response curves), relating a population's *treatment* or *exposure* to the observed *effect*

The strength of association may also be measured by "probabilistic dependence". In comparison to rate of growth and the dose-response gradient, this may be a less informative indicator, in that it need not provide any information as to the functional form of the relationship between cause and effect, but merely denotes the presence of an asymmetry in the conditional distribution of the two variables.

In our system we distinguish between "rate of growth", "dose-response relationship" and "probabilistic dependence" as different indicators of causality corresponding to Hill's "strength of association".

We also introduce "difference making" as a *perfect* indicator of causality: this is to be inferred either through experiment or through "intervention" as intended in the causal graphs literature (Pearl 2000; Woodward 2003) (see Section 5.2).

The main distinction between the three former indicators and the latter one is that difference making represents an asymmetric relationship (it is about what makes a difference to what), whereas the various forms of strength of associations are all symmetric in principle, and therefore cannot provide straightforward information about the direction of the relationship. This also follows from the very same reason why dose-response relationship, rate of growth and probabilistic dependency are only imperfect indicators of causality, whereas difference making is a perfect one (more on this below in Section 3.2.6, Page 19, and Section 5.2).

3.2.2 Consistency

Hill's second viewpoint "consistency", also refers to stability in that it centers on the question: "Has [the association] been *repeatedly* observed by different persons, in different places, circumstances and times?" (emphasis added). However, by referring to *repeated* observations, consistency also relates to replication of studies with identical methods, or in (systematically) varied study settings.

Indeed, as for any other empirical science, clinical trials and epidemiological studies cannot test a given hypothesis but in highly contingent study settings, together with its theoretical/methodological assumptions, and *ceteris paribus* clauses. Hence, systematic variation of study design and setting serves to provide evidence that the result is not an artifact of the particular circumstances in which a given study has been carried out, or of the particular method, or theoretical model adopted, and related assumptions (see also Kuorikoski et al. 2010). Hence the role of consistency should be distinguished along the following lines (see also Table 1):

1. Replication of (ideally) identical studies (same "background conditions" – same inclusion and exclusion criteria, mode of administration/exposure, etc. – and same design, e.g., cohort study, RCT, etc.): this is a means to increase accuracy of measurement.
2. Replication of the observation through different methods, but analogous background conditions: this should test the results against the suspicion of being created by the specific study design/setting ("study artifacts") and guarantee "methodological robustness".
3. Replication of the observation through similar methods, but in different background conditions: this should test the stability of the causal link itself in different populations/circumstances and show the extent to which it is "ontologically robust" (Wimsatt 2012; Open Science Collaboration 2015; Meehl 1990; Woodward 2006).

Table 1 Different types of epistemic gain from different types of study replication

Replication of studies testing the same causal link . . .	increases confidence in . . .
in ideally identical conditions (same method, same background conditions)	accuracy of measurement
with different methods under same background conditions	methodological robustness
with the same method under different background conditions	stability of the causal link across different populations

Meehl (1990) represents the logical structure of theory testing as follows:

$$(T, A_t, C_p, A_i, C_n) \rightarrow (O_1 \supset O_2) \ .$$

Where T represents the theory to be tested, A_t its ancillary assumption, C_p denotes the *ceteris paribus* clauses, A_i the methodological assumption of the specific study design used, and C_n the specific and absolutely contingent conditions of a given individual experiment. The arrow denotes entailment and the horseshoe $\supset$ between O_1 and O_2 material entailment. So, the left-hand conjunction is falsified, *modus tollens*, if you observe $(O_1, \neg O_2)$, instead, observing (O_1, O_2) provides inductive support to it.

Study replication of the first kind (1) serves the purpose of verifying whether the results, positive or negative, are due to the contingent settings of the individual study (C_n); systematic variation of experimental/observational settings (2) is meant to identify the influence of the methodological settings (A_i). Replication (3) allows one to test the possible violation of *ceteris paribus* clauses (C_p); theory-related ancillary assumptions (A_t), and any testable consequence of the theory itself (T). Hence, the confirmatory role of replication consists in enhancing the probability of the theory T by decreasing the probability that the evidence provided by the study for it is instead due to A_t, C_p, A_i, or C_n.

In the specific context of causal inference by means of clinical or epidemiolgical studies, the role of replication and systematic study variation can analogously be related to the various components of scientific inference. So, T may represent the hypothesis of the causal connection itself; A_t the ancillary assumptions, C_P, the "everything else being equal" clause in relation to the investigated drug, that is, the possible interacting factors such as age, sex, co-morbidity etc., A_i represents the methodological assumptions, that is those aspects of the study design which make a difference to the result and its interpretation with respect to other kinds of designs (e.g., a cohort study with respect to a case-control study); and C_n represents the contingent circumstances and conditions of any given study.

Consequently, from a confirmatory point of view, consistency across studies may mean very different things depending on whether such studies share the same design, the same kind of population or both.

In our framework, consideration of the confirmatory value of consistency is incorporated in the general structure of the network, where multiple reports for the same indicator may be related by the same reliability or relevance nodes if they share the same methodology or background conditions respectively.

3.2.3 Specificity: quantitative and qualitative versions

Specificity also refers to diverse phenomena. The traditional concept of specificity (following Hill) approximates the classic conception of a cause as a *necessary and sufficient* condition for its effect to occur, in contrast to the possibility of it being produced by other candidate causes. In particular, specificity is interesting when considered as an (ideally) bijective function holding between sets of cause classes and sets of effect classes; for instance when various kinds of a class of toxic agents are related to various kinds of cancers through biunivocal relationships.

Specificity as a property of causality has been discussed by Lewis (2000), Waters (2007), and Woodward (2010) as the sort of functional relationship which systematically holds between the values of a variable and the values of another variable, such that changing the value of the former in specific ways also changes the values of the latter in specific ways, ideally in a bijective fashion. Lewis uses for this kind of causal relationship the term "influence" and describes it as follows: "*C will influence E to the extent that by varying the state of C [...] we can vary the state of E in a fine-grained way*" (Lewis 2000, p. 305).[7] In itself, an ideal bi-conditional association with no residual instantiations of Cs not accompanied by Es, or vice versa, would not guarantee causation: specificity is not an exclusive property of causation (also conventional codes are specific in this sense, see Osimani 2014a; Kment 2010, p. 83). However, it is very unlikely that such bi-conditional relationships can occur by chance alone, therefore, specificity may be considered as an indicator of a causal association being present, in that it manifests an underlying robust relationship, which explains the systematic correspondence between the two relata.

Another, related notion of specificity refers to the "geometrical complementarity" on which many biological phenomena are based; such as the key-hole relationship between antigens and antibodies, or between target receptors and drug molecules (see Weber 2006). This kind of specificity is relevant to causal assessment and prediction in pharmacology for methods (such as so called "systems pharmacology"), that try to infer the possible effects of drugs by identifying the sets of possible families of receptors – and related proteins – which a certain drug molecule could bind to on grounds of its "affinity" with them. Indeed, affinity is a function of stereometric properties (structural and biological similarity); see for instance Xie et al. (2009). This kind of specificity pertains to mechanistic reasoning, hence it will be considered insofar as it is used to glean causal evidence from knowledge about mechanisms (see below, biological plausibility: evidence of mechanisms, Section 3.2.6).

Specificity and stability are independent concepts: a causal relationship may be at the same time highly specific and unstable (e.g. multifactorial genetic diseases), or it may be highly stable but show a low degree of specificity (e.g. intestinal inflammation caused by various kinds of antibiotics).

Neither stability nor specificity will be used in our framework to distinguish different types of causes, but rather as signs of the presence of a causal connection. In our framework, specificity is encoded as *difference-making* since those very studies aiming at detecting causal efficacy of a drug under investigation also yield information about the variance of the effect under different tests.[8]

[7] Woodward (2010) uses specificity to distinguish between different kinds of causes, thereby leaving room for ontological pluralism, and also allows for specificity (as well as stability) to come in degrees.

[8] Since specificity-as-bijection (referring to a property of the investigated nexus between cause and effect) supports the causal hypothesis $D©H$ by *excluding* alternative explanations of the observed effect (and ideally also alternative effects of the tested drug), we propose to model these alternatives on the same categorical level as our main hypothesis, with the same methodological arsenal for testing and confirming (or rejecting) them. It is the confirmation of $D'©H'$ together with the rejection of $D'©H$ and $D©H'$ that makes our actual hypothesis a specific relation, thereby lending additional confirmatory support to it.

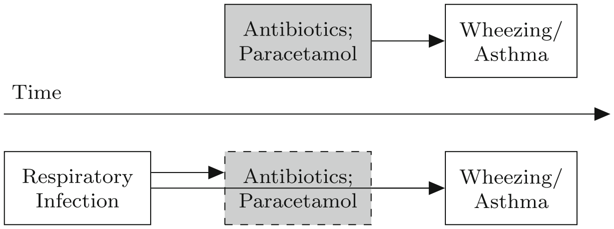


Fig. 4 *Confounding by indication* in the case of the association between antibiotics and asthma (cf. Heintze 2013, p. 1205, Figure 4). Where the upper structure contains a direct causal link, the second structure includes a third and earlier factor, respiratory infection, triggering both the prescription of drugs and what turns into wheezing or asthma at a later time

3.2.4 Temporal order, distance, and duration

In his question "which is the cart and which the horse?", Hill addresses the aspect of temporal precedence, seen as one of the most important markers of causality. The importance of this criterion is mirrored by the fact that many theorists of causality consider it a necessary prerequisite, to postulate alignment of the causal and the temporal direction, or even explicitly incorporate it in their formal framework.[9] For instance, Suppes (1970) goes beyond Reichenbach's *common cause principle* in explicitly building on the direction of time in his probabilistic definition of causality: an event genuinely causes a subsequent second event if it is identified as a "prima facie cause" – i.e., it *precedes* the effect and raises its probability – and guaranteed not to be an instance of "spurious causation". Hence, although temporal precedence is a necessary condition for causality, it is not sufficient for it because of the possibility of confounding.

The methodological literature speaks about "reverse causation" in cases where precedence of time, together with statistical association, might give the false impression that the preceding phenomenon causes the succeeding one, whereas the contrary holds. In epidemiology, this is mainly due to issues related to *duration* and *manifestation* in time of causal phenomena.[10]

For example, the recent debate around the causal association between paracetamol and asthma centers around the possibility that cohort studies showing a statistical association between the drug and the disease, may not warrant a causal conclusion, notwithstanding the temporal precedence of paracetamol consumption with respect to disease inception, because subjects affected by asthma in its subclinical phase (i.e., when it has not been diagnosed yet) have a higher than average tendency to develop colds, fever or rhynitis, and therefore to take more paracetamol than the unaffected population (Heintze and Petersen 2013; Weatherall et al. 2014). Figure 4,

[9] Precedence in time is considered so essential to causes that Russel bases his denial of their existence on the temporal symmetry of laws in physics, (Russell 1912).

[10] From a more general perspective this precisely touches upon the difficulty of defining or describing an event, as discussed, e.g., by David Lewis in "Counterfactual Dependence and Time's Arrow" (1979) and "Events" (1986).

replicated from Heintze and Petersen (2013), shows the potential common cause structure explaining the association between the use of antibiotics or paracetamol and later wheezing or bronchial asthma.

In our framework, temporality will be expressed straightforwardly by a *temporality* variable to sum up the above distinctive criteria.

3.2.5 Biological gradients and dose-response models

Related to specificity as influence is also the dose-response relationship, where a clear pattern of quantitative dependency manifests. Quite in line with Lewis' idea of fine-grained relevance as a characteristic feature of causation, Hill sees it as a strong indicator of a causal relation if an investigation reveals a biological gradient, in that a clear dose-response curve admits a simple explanation.

In analogy to the strength of association indicator, also the biological gradients splits up into the three dimensions "rate of growth", "dose-response relationship", and "probabilistic dependence".

3.2.6 Inferential patterns

Plausibility of the Biological Mechanism

The role of evidence about possible/plausible/actual mechanisms (Machamer et al. 2000; Darden 2006; Craver 2007) linking the putative cause to its phenotypic effect is strongly debated in philosophy, especially in relation to evidence standards. Philosophers closer to the Evidence Based Medicine approach, even in recognising some value to knowledge about mechanisms, still doubt that they can complement statistical black-box evidence because of the limited and fragmentary knowledge of the "causal web" in which they are embedded (see for instance Howick 2011). Other philosophers instead generally recognise that knowledge about mechanisms plays a plurality of roles both in combination with statistical information and in a stand-alone fashion:

1. Following the philosophical analysis of causal explanation, the most traditional epistemic role assigned to knowledge about mechanisms is to provide the ontological rationale for observed regularities (Salmon 1984).
2. Knowledge about mechanisms can also constitute a sort of double check for causality (Salmon 1997; Russo and Williamson 2007; Clarke et al. 2014).
3. Mechanisms have a methodological relevance in that they are supposed to provide the basis for extrapolation (Cartwright 2007b; Luján et al. 2016), and are important for supporting the reliability of model assumptions as well as for interpreting experiment results (for instance a two-way curvilinear causal interaction cannot be detected or may be misinterpreted by linear regression models).
4. Finally, mechanisms are held to have an epistemological/theoretical relevance, in that they can provide the hypothesis which puts together disparate data (through abductive inference). In this sense they provide the basis for the accumulation of knowledge and scientific progress (see also Craver 2007).

In the epidemiological literature, especially in narrative reviews, it is typical to combine evidence about mechanisms with statistical evidence at the population level, as an argumentative move in favour of causal hypotheses which are only suggested by statistical associations. For instance, going back to our debate about the possible causal association between paracetamol and asthma, a series of reviews (Shaheen et al. 2000, 2008; Katayoun 2005; Henderson and Shaheen 2013; Allmers et al. 2009; McBride 2011; Heintze 2013; Martinez-Gimeno and García-Marcos 2013) present both population-level studies as well as evidence of molecular and cellular mechanisms from bio-essay or animal studies showing plausible biological pathways leading from consumption of paracetamol to the development of hyper-responsive reactions and asthma, see Osimani (2014b). This latter evidence is meant to provide support for the "physical" connection, if the observed statistical association was due to a causal association.[11] In our framework this dimension will be mapped onto the indicator for mechanisms.

The plausibility part of the "plausibility of mechanisms" indicator refers to the general fit of the hypothesised mechanisms to available background knowledge and this leads us directly to the subsequent "viewpoint" on causality (as Hill would call it), namely: "coherence of evidence". Neither coherence of evidence nor the subsequent viewpoints listed by Hill ("experiment" and "analogy") are strictly speaking indicators of the presence of causal relationships themselves. Rather, they refer to the inferential/methodological process itself, and they appeal to particular methods (experimental), or kinds of reasoning ("analogy"), or theoretical/epistemological virtues ("coherence") which may be adopted to "optimise causal inference".

Coherent evidence

Coherence is a property of the body of evidence, rather than of the phenomenon under investigation (here, the causal link between drug and side effect). Hence, coherence may involve the concept of consistency seen above and therefore denote 1) a property of a set of measurements, related to the same investigated *parameter*, both in the sense that they all indicate a positive (or a negative) effect, and in the sense that the strength of the effect size does not exceed statistical variability across studies – this sort of coherence is generally referred to as consistency of results, and is established through replication; 2) a property of a set of various pieces of evidence related to the

[11] The role of evidence about mechanisms of chemical substances in risk assessment has been recently analysed by Luján et al. (2016). Two issues are particularly relevant for the present purposes: 1) the questioned applicability of animal data in humans; 2) the lack of guarantee that similarity of modes of action may warrant extrapolation of phenotypic effects from one chemical to another. Both issues relate to the problem of extrapolation: the former regard whether a given chemical will produce the same effect in the study and in the target population; the latter refers to whether similar chemicals produce similar effects (on a given population). In our framework the problem of extrapolation is addressed with the following in mind: I. As explained in Section 2.3, in the case of risk assessment the main concern relates to false negatives; hence any signal should be accounted for as a possible sign unveiling latent risks – "If it happened there, it can also happen here"; II. Warrant for extrapolation is also taken to come in degrees and therefore is incorporated in a probabilistic approach. This lets the degree of confidence in such warrant guide the decision at hand in combination with other relevant dimensions (as illustrated in Section 2.2).

same testable consequence of the hypothesis, investigated through diverse kinds of *methods*: methodological robustness or “triangulation” (see Wimsatt 1981; Wimsatt et al. 2012); 3) a property of a set of pieces of evidence related to diverse testable consequences of the investigated *hypothesis*. This latter sense is the less investigated and accounted for in the methodological literature.

In the epidemiological literature, consistency and coherence are not explicitly distinguished and evaluation of the latter is left to informal/implicit judgment in narrative reviews. Instead, the philosophical literature has investigated coherence in several respects: 1) general epistemology offers “coherentism” as a response to skepticism in alternative to “foundationalism”: according to this view beliefs are justified by their fitting together in a system, and standing in a relation of mutual support (BonJour 2010), like the stones of an arc (*simul stabunt, simul cadent*); 2) formal epistemology has investigated the confirmatory value of coherence of beliefs, also by developing various measures of coherence, in the attempt both to formalise its content and to track its truth-conduciveness; see for example Crupi and Tentori (2014), Fitelson (2003), Bovens and Hartmann (2003), Dietrich and Moretti (2005), and Moretti (2007).

As a theoretical virtue, coherence is particularly relevant in a context, such as risk assessment, where evidence may come from different sources and relate to diverse levels/dimensions of the suspected causal association between drug and side-effects.

We adopt and adapt such accounts of coherence in order to develop a system of evidence amalgamation which naturally incorporates coherence as confirmatory, by virtue of the way the system lets heterogeneous pieces of evidence interact and thereby jointly contribute to the (dis-)confirmation of the investigated hypothesis.

Support by experiment

The scientific method strongly relies on systematic observation and experiment. In particular, carefully controlled experiments are considered a privileged way to inquire nature, in that they ideally allow the scientist to isolate the phenomenon under investigation from interfering and disturbance factors. More specifically, the systematic variation of experimental conditions recalls Mills method of eliminative induction (Mill 1884), in that it allows to see the behaviour of the studied phenomenon in combinatorial rearrangement of different circumstances, encoding both methodological and theoretical assumptions (different “possible worlds”).

However, traditional experiments in physics are very different from experiments in biology, pharmacology, and medicine (but also sociology and psychology). In physics, the experiment is meant to test a theory by comparing the observed value and the value predicted by the theory: statistical significance speaks directly *for* the theory in that it reflects a small discordance between the two values. In the latter sciences instead, the tendency has prevailed to test the null hypothesis, that is, the catch-all complement of the claim that the experiment is supposed to provide evidence for (see also Meehl 1990).

This has raised various criticisms both addressed against the methodology of hypothesis testing itself (see for instance Howson and Urbach 2006) as well as concerning the epistemic value of randomised controlled trials in medicine (Cartwright

2007a; Papineau 1993; Worrall 2007b, 2010; Teira 2011; La Caze et al. 2012). In the latter case, the charge involves the alleged epistemic virtues of randomisation:[12] the exclusive reliance on RCTs for the justification of causal claims is considered to be wrongheaded for various reasons; both related to the kind of information which they provide, and to the kind of information which they are not able to incorporate or account for.

Indeed, whereas classical experiments in physics allow the scientist to observe the behaviour of the investigated phenomenon in an array of different "possible worlds" (in different *scenarios* or under different *initial or boundary conditions*), and compare systematic differences among such situations, randomisation is blind to such specific settings. Its outcome is rather to neutralise their effects on the final result, by creating two populations, one for the treatment and one for the control group, where the same "worlds" should be represented in the same proportion. This should guarantee that the different results possibly observed at the end of the trial are due to the treatment and only to it.

Hence randomisation, as a means to isolate the putative cause from other possible confounding factors, loses much information with regard to the specification of possibly relevant mediating and interacting causes. However, it provides a much higher guarantee with respect to confounding than observational studies do. Ceteris paribus, RCTs fare better in distinguishing spurious from genuine causes.

Since Rubin (1974), the standard conceptualisation of causal claims resulting from RCTs (and comparative studies) is counterfactual: the "causal effect" is the difference between what would have happened to the subject, had it been exposed to the treatment and what would have happened to it, had it been exposed to the control. Since the subject cannot undergo the same experimental conditions at the same time, the causal effect is calculated as the average difference of the effects observed in the group of exposed and the group of unexposed subjects.[13]

Causality has indeed been analysed in terms of counterfactuals in several respects. Lewis (1973, 1986, 2000) proposes a possible-worlds semantics for the truth conditions of individual causal claims in terms of counterfactual dependence; Woodward (2003) identifies necessary and sufficient conditions for causality in terms of invariance under intervention, where his notion of intervention captures the counterfactual

[12] Randomisation has putatively two main roles: 1) in the long run, it should allow the investigator to approach the true mean difference between treatment and control group; however it is unclear what this true underlying population probability denotes when we are dealing not with population of molecules for instance, but with population of patients undergoing medical interventions, where heterogeneity among individuals can at most allow for an aggregate average measure. Furthermore, it is obviously unethical and unfeasible to re-sample the same subjects of an experiment again and again, and even if this were possible, the subjects who were administered the drug in the first round would undergo physiological change; consequently, the successive trial population would no longer be the "same" (Worrall 2007a); 2) randomisation (together with intervention and blinding) should guarantee the internal validity of the study by severing any common cause, or common effect, between the investigated treatment and its putative effects (i.e., avoidance of confounders and (self-) selection bias). This kind of objective is supposed to justify the primary role assigned to randomised evidence by so called evidence hierarchies, see Section 6 below.

[13] This is also known as the "potential outcome approach" to causal inference.

gist of causal graphs and structural equation modelling. Finally, Pearl (2000) is focused on counterfactuals related to potential effects of interventions, and relies on causal knowledge to predict the effect of such interventions (e.g., policy interventions). Hence, counterfactuals have different roles in analysing causal claims, defining causality or using causal knowledge for predicting the effect of interventions. However, they share the intuition that a cause must make some difference to whether the effects occur or not (holding other variables fixed).

In our framework we identify RCTs as particularly reliable sources of evidence for the difference-making effect attributed to causes by the philosophical literature; and we understand difference-making as *ideal controlled variance* along the concept of *intervention* in manipulationist theories of causation. For example, Woodward (2003) carefully characterises intervention variables for the purpose of testing the hypothesised causal connection between two events X and Y. Figure 5 illustrates the idea: If the unknown underlying causal connections form a fork (first graph in Fig. 5), the introduction of a suitably chosen intervention variable I will lift X from the influence of X and Y's common cause and refute the hypothesis that X is a cause of Y (second graph in Fig. 5). Nevertheless, if I is not chosen well it might indeed lift X from its parents' influence but introduce a new dependency between X and Y, leading to the false conclusion that X causally influences Y (third graph in Fig. 5). In contrast to Woodward, Pearl (2000) expresses interventions as local surgeries in the causal model, subjecting X to ideal (external) control by directly setting its value (qua $do(x)$), thereby abstracting from the possibility of finding a suitable intervention variable. X is then 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). We see our concept of difference-making as closely related to Pearl's definition of what it means to be a cause, and accordingly, difference-making will be a *perfect* indicator of causation (also see below, Section 5.2).

Support by analogy

Hill briefly mentions reasoning by analogy as additionally contributing to the assessment of the causal claim: if a specific drug is on trial, available evidence of a similar

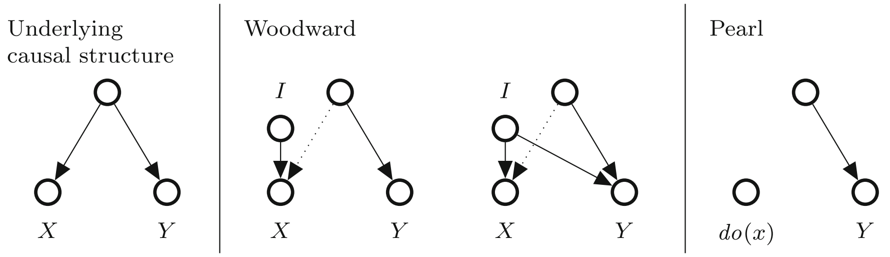


Fig. 5 Interventions can be used to test the underlying causal structure and learn more about the relation between event X and Y. Woodward's interventions require an additional, suitably defined intervention variable I (overriding causal influence on X, marked by *dotted edges*), whereas Pearl's interventions are expressed as local surgeries in the causal graph

second drug's effects might be used for inference about the former. This touches upon central and notorious epistemological questions: What does it mean to be sufficiently similar in the case under consideration? In what way does the difference between the first and the second drug influence changes in expected outcome values? How specific are a drug's properties? If they are highly specific – to what extent can this drug be used in an analogical argument, if at all? Although similarity seems to be a concept difficult to spell out in formal terms,[14] the applicability and fruitfulness of *parallel reasoning* is of great interest (see Bartha 2010 and Hesse 1959), and analogical arguments are employed across disciplines.

Physicists, e.g., transfer abstract structures of analogue reasoning to analogue simulations with which physical systems (presumedly similar in all relevant aspects) are tested under syntactic isomorphism (see, e.g., Unruh 2008; Dardashti et al. 2016; Hesse 1952 for discussions of analogue arguments in physics). Scientific discovery is oftentimes propelled by analogy, as, e.g., in the nineteenth century, when secured knowledge about acoustics was employed in the discovery of spectral lines. Guided by the image of a harmonic oscillator, physicists were able to focus their attention to groups of spectral lines with specific frequency patterns from the beginning (see Bartha's in-depth overview of analogical arguments in Bartha 2013). When coupled with a suitable theory of confirmation, analogy can finally be used to support a scientific hypothesis where only evidence from an analogue system is obtained (see, e.g., Hesse 1964 or Poellinger and Beebe 2017). This is of special interest for our purposes in this paper.

In pharmacology and epidemiology, explanation and prediction *by analogy* rest both on sufficiently well-described background conditions and knowledge about the relevant biological mechanisms at work. Describing all relevant differences between two drugs might be the first step towards justifying assessment by analogy – the second step might then be inference in a unified model where all the relevant differences are integrated as parameters. Formal models relating different pieces of evidence can precisely be of help for this task. Once relevant influences are distinguished from irrelevant ones and the contribution of differences in the relevant factors are determined, analogy will *justifiedly* help in identifying causation.

Reasoning by analogy is also at the basis of inductive inferences from study to target population. Indeed, because of the context sensitivity of many causal associations in the biological realm, these can hold only in specific populations, and therefore evidence about causal effects related to one population may not license similar conclusions about another population, unless the two population are *analogous*.

Our framework can be operationalised for the assessment of support by analogy in two ways:

[14] David Lewis' bases his formal account of causation implicitly on the concept of comparative similarity and concedes the following: *"We do make judgments of comparative overall similarity – of people, for instance – by balancing off many respects of similarity and difference. Often our mutual expectations about the weighting factors are definite and accurate enough to permit communication. [. . .] But the vagueness of over-all similarity will not be entirely resolved. Nor should it be. The vagueness of similarity does infect causation, and no correct analysis can deny it."* (cf. Lewis 1973, pp. 559–560)

1. The investigated causal hypothesis $D©H$ (with all of its sub-structure) may be related to a second causal hypothesis $D'©H'$ which has already been confirmed in specific previous studies. Now, if scientists have sufficient reason to propose an analogy relation between D and D' (e.g., a high degree of chemical or functional similarity), knowledge about this second causal hypothesis supports the first hypothesis $D©H$ *via analogy* ("horizontally" on the same level of investigation).
2. Since in general no piece of evidence comes from an experiment or a study conducted on the target population, the question of applicability of the study's findings must be phrased in terms of the similarity between study and target populations: If study and target populations are sufficiently similar, researchers are licensed to reason about causal links in the target population by analogy with their test cases. The degree to which this kind of transfer is licensed is encoded in our framework as an attribute of available reports (see Section 3.3 below for a discussion of the relevance of evidence and how it "vertically" influences the assessment of the causal hypothesis).

Relativising the justification of analogue reasoning to a (comparative or numerical) distance measure (i.e., the similarity between drugs or populations) might of course be criticised due to the perspectival nature of *similarity* and *relevance*. Nevertheless, once this measure is agreed to be of sufficient strength, our framework can be used to express the confirmatory dynamics of support by analogy.

3.3 Relevance and reliability

In the literature on (the philosophy or even science of) evidence, we find the idea that studying the concept of "dimensions" of items of evidence can help address the notoriously tricky notion of "weight of evidence" (Weed 2005). Schum put forward a substance-blind classification of inferential power of evidence employing the dimensions: "credibility" and "relevance", see Schum (2011). To score an item of evidence on these dimensions one needs to answer the following questions "*Can one believe the reported evidence?*" and "*How does an item of evidence bear upon the proposition of interest?*".

In the same line, Roush argues that good evidence for the user has to be "credible" and "relevant" (see Roush 2005, Chapter 5).

In Cartwright (2008) and Cartwright and Stegenga (2011), Cartwright and Stegenga shift the focus to the user of evidence (policy or decision maker). With this move they make room for a more pragmatic notion of relevance which refers not only to the confirmatory force of the evidence with respect to the hypothesis under investigation, but also to whether the evidence acquired in the study may license the same conclusion in domains/populations which may differ from the study sample (external validity, extrapolation). We clearly demarcate these distinct notions of relevance by adopting a "relevance dimension", which refers to external validity issues, as a notion separate from the standard Bayesian relevance (as paradigmatically measured by probabilistic dependence). Also we distinguish these from the reliability dimension.

Relevance Ideally, pharmacological studies would license the same inferences for the studied population and the target population of future drug users. In reality, studies are not conducted on the *entire population of future drug users* but on a much smaller number of patients, see Chan and Altman (2005, p. 1180), Button et al. (2013) for this problem in neuro science, Doll and Peto (1980) in cancer research and a philosophical discussion of this problem in Worrall (2007a, p. 992). Additionally, studied populations, in particular in RCTs, often fail to be representative for the target population due to strict patient inclusion criteria, see Revicki and Frank (1999) and Upshur (1995, p. 483). Therefore, there is a need to reason by analogy from the studied population to the population of interest (see Section 3.2.6).[15] The relevance pertaining to an item of evidence measures how well the observed results in a study population can be transferred to the target population of future drug users.

Reliability The "credibility" dimension of evidence relates to the source which originates it and the way it has been collected. Bovens & Hartmann, see (Bovens and Hartmann 2003, Chapter 4), elaborate on "reliability" as an instrument for Bayesian hypothesis confirmation where it is construed as i) a function of an instrument's accuracy or ii) the credibility of testimony.

The sums at stake in drug licensing decision problems are enormous (revenues of successful drugs can reach billions of dollars within a few years) it is hence not surprising that vested interests may influence the flow of information. A point in case for vested interests to playing an inglorious role is the story of "Vioxx", see Carné and Cruz (2005), Horton (2004), Jüni et al. (2004), Krumholz et al. (2007), and Holman and Bruner (2015). The abstract worry that financial conflicts of interests lead to a "bias in the synthesis and interpretation of scientific evidence" seems to be not merely abstract, see Bes-Rastrollo et al. (2013) and Dunn et al. (2014). A low reliability of an item of evidence may be due to a number of methodological flaws (confounding, biased studies, more broadly: poor design of a study, sub-optimal data recording) as well as not fully objective sources.

In our framework we shall use *variables for every piece of evidence*. To every such variable two further variables pertain, a *variable for the relevance* and a *variable for the reliability*.

4 Modeling scientific inference in Bayesian epistemology

4.1 The Bayesian paradigm

Bayesian epistemology is a current paradigm in the philosophy of science, see for example Howson and Urbach (2006) and Weisberg (2015). It has two virtues which are key for us. Firstly, it complies with Carnap's principle of total evidence

[15] One particular inference by analogy is that from animal studies/models to a human target population. In LaFollette and Shanks (1995), it has been argued that animal studies are only good for hypothesis discovery. We side with Baetu (2016) in thinking that animal studies are one important piece to the puzzle to predicting drug reactions.

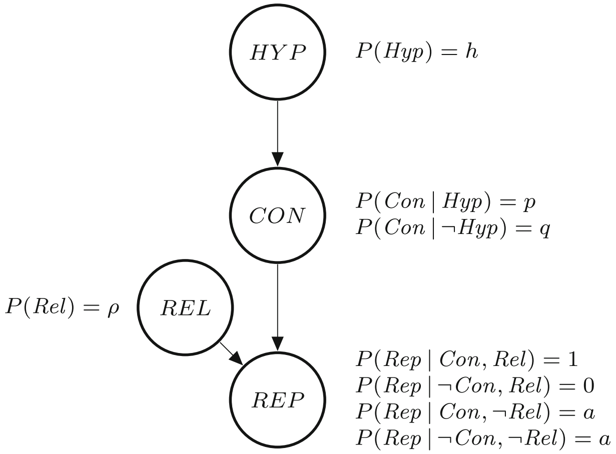

HYP
$P(Hyp) = h$
CON
$P(Con \mid Hyp) = p$
$P(Con \mid \neg Hyp) = q$
$P(Rel) = \rho$
REL
REP
$P(Rep \mid Con, Rel) = 1$
$P(Rep \mid \neg Con, Rel) = 0$
$P(Rep \mid Con, \neg Rel) = a$
$P(Rep \mid \neg Con, \neg Rel) = a$

Fig. 6 Hypothesis testing in the Bayesian framework of Bovens & Hartmann for one single testable consequence. All necessary parameters of the prior probability are displayed in terms of lower-case letters

(Carnap 1947) which demands that *all available evidence* is taken into account when determining probabilities. This is achieved by updating prior probabilities by the evidence propositions yielding posterior probabilities.

Secondly, it allows for a nuanced assessment of hypotheses after incorporating all available evidence into the posterior. The degree to which the evidence boosts (or lowers) the probability of a hypothesis is the (dis-)confirmation the evidence provides to the hypothesis.[16] In Bayesian epistemology, the role of evidence in (dis-)confirming a hypothesis is thus grounded on a sound methodological bases.

A theoretical framework which allows us to distinguish the different epistemic levels of the problem at hand, as well as their interaction is presented in Bovens and Hartmann 2003, Chapter 4). This framework is a model for epistemic dynamics underpinning scientific inference and it provides a mathematical explanation for it by labeling nodes with epistemic categories. In Section 5, we will adapt their work for our purposes of causal inference in pharmacology.

4.2 Scientific hypothesis confirmation according to Bovens & Hartmann

The Bayesian network model of Bovens & Hartmann, presented in (Bovens and Hartmann 2003), (see (Darwiche 2009; Neapolitan 2003) for introductions to Bayesian networks) consists of the hypothesis, (some of) its observable consequences, reports on whether theses consequences were born out in experiments and the reliability of instruments used in these experiments. The graph structure of the

[16] The reader is referred to the very recent (Crupi and Tentori 2014) which discusses two leading Bayesian confirmation measures in detail.

Bayesian network represents conditional independencies between the variables. Conditional probabilities attach to every variable which specify the probability of a variable given its parents.

The following binary propositional variables are used: A variable HYP where the intended meaning for Hyp is that "the hypothesis is true", similarly for variables Con_i ("consequence i holds"), Rep_i ("consequence i is reported")[17] and Rel_i ("report i is reliable"), cf. (Bovens and Hartmann 2003, p. 89). According to the Bayesian paradigm, a prior probability function P, defined over the algebra generated by these variables, is selected. Naturally, P is constrained to respect the conditional independencies encoded by the graph $\mathcal{G}$. Updating the prior P, by conditionalising, then allows Bovens & Hartmann to calculate posterior probabilities given experimental results.

The set of meaningful conditional probabilistic independencies in prior P can be read off by means of the graphical d-separation criterion (Pearl 2000). The graph $\mathcal{G}$ in Fig. 6 depicts the situation for one single consequence. The general case for more consequences is more involved and depicted in Fig. 7. These conditional independencies – denoted by $\perp\!\!\!\perp$ – are [18]

$$HYP \perp\!\!\!\perp REL_i \quad \text{for all } i \tag{1}$$

$$CON_i \perp\!\!\!\perp REL_i \,|\, HYP \quad \text{for all } i \tag{2}$$

$$REP_i \perp\!\!\!\perp HYP \,|\, REL_i, CON_i \quad \text{for all } i \tag{3}$$

$$\{CON_i,\ REL_i\ REP_i\} \perp\!\!\!\perp \bigcup_{k\neq i}\{CON_k,\ REL_k,\ REP_k\} \,|\, HYP. \tag{4}$$

The choice of prior is further constrained by

$$P(Con_i \,|\, Hyp) = p_i \; > \; q_i = P(Con_i \,|\, \neg Hyp) \tag{5}$$

$$P(Rep_i \,|\, Con_i, \neg Rel_i) \;=\; P(Rep_i \,|\, \neg Con_i, \neg Rel_i) = a_i \tag{6}$$

$$P(Rep_i \,|\, Con_i, Rel_i) \;=\; 1 \tag{7}$$

$$P(Rep_i \,|\, \neg Con_i, Rel_i) \;=\; 0. \tag{8}$$

Bovens & Hartmann take Eq. 5 to be their definition of what it means to be an observable consequence of a given hypothesis, see Bovens and Hartmann (2003, p. 90).

Equation 6 refers to the convention that when an instrument is unreliable, then the probability of receiving a report does not depend on whether the consequence holds. When the instrument is fully reliable, then the probability of receiving a report that the consequence has been observed equals one, if the consequence holds; see Eq. 7. Vice versa, a fully reliable instrument produces a positive report with probability zero, if the consequence does not hold, see Eq. 8.

They then determine the (posterior) probability of the hypothesis being true, given a report and its reliability (Bovens and Hartmann 2003, p. 92, Equation 4.5). This

[17] $\neg Rep_i$ means that "not consequence i is reported" rather than "consequence i is not reported".

[18] Where REL_i and CON_i denote the parent variables of REP_i.

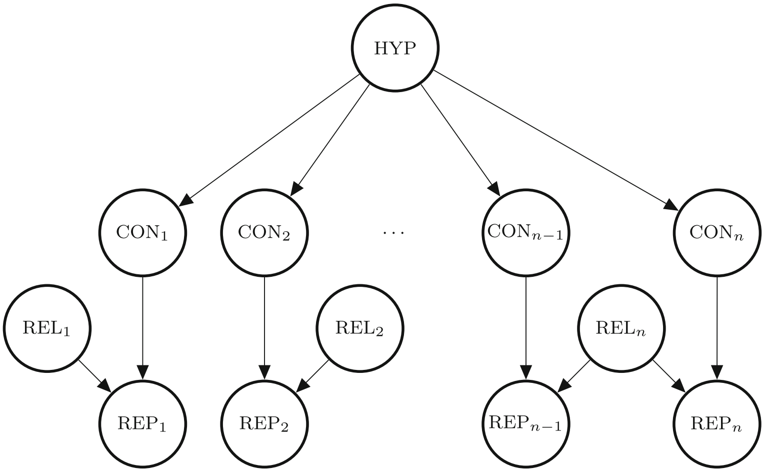


Fig. 7 Evaluating a hypothesis with multiple testable consequences according to Bovens & Hartmann. A reliability node with multiple children is used for instruments which were used in multiple experiments

probability can be computed directly from the conditional probabilities specified at the nodes in the Bayesian network. The comparison of prior and posterior probability is at the heart of Bayesian confirmation theory.

The rationale for arranging the theoretical layer (HYP, CON) and the empirical/methodological layer (REP, REL) in the proposed way is summarised in the following:

1. Testable consequences of the hypothesis are inserted as intermediate nodes between the HYP-node and the report nodes: $HYP \perp\!\!\!\perp REP_i \,|\, CON_i$ for all i. According to Bovens & Hartmann, this formally captures the fact that the hypothesis cannot be tested directly, only observable consequences of the hypothesis are testable (cf. Bovens and Hartmann 2003, p. 89).
2. All consequence nodes (together with their respective descendants) are conditionally independent given HYP: $REP_i \perp\!\!\!\perp REP_j \,|\, HYP$ for all i, j with $i \neq j$. This models the situation where a range of consequences can be assessed by multiple independent tests (see the left portion of Fig. 7 and cf. Bovens and Hartmann 2003, p. 98). If, however, the dependence of studies or tests is to be marked in the structure, one REL node might also have more than one REP node as a child, thereby creating dependency between reports (see the right portion of Fig. 7).
3. All *Rel* nodes are unconditionally independent of their non-descendants. That means: $REL_i \perp\!\!\!\perp nonDesc(REL_i)$ for all i; i.e., in the case of all reports being independent we have $REL_i \perp\!\!\!\perp HYP$ for all i (with i indicating a report for the i-th consequence). This independence formalises the assumption that from learning something about the reliability of a report we cannot infer anything about the truth (or falsity) of the hypothesis (cf. Bovens and Hartmann 2003, p. 58).

4.3 Virtues of the Bovens & Hartmann approach

Probabilistic networks have found wide-spread use in applications in a great number of domains including medicine (e.g., differential diagnosis). These networks implement knowledge from a very specific domain and nodes generally represent concrete phenomena or entities of interest.

Bovens & Hartmann draw on this formal apparatus to abstractly model scientific inference and to formally relate higher level epistemic categories. While the mathematics do not change, the graph provides an illustration of the epistemic dimensions at stake and thereby provides greater insight into some methodological issues by offering, as it were, a mathematical explanation of their dynamics. Indeed, this sort of representation allows us to single out in the mathematical formulae the specific role played by each epistemic dimension in the inferential dynamics; e.g., the role of reliability with respect to the propagation of confirmation in connection with replication of studies and with heterogenous sources of evidence (see Section 3.2.2).

5 A framework for the assessment of harms in pharmacology, part 2: structure

5.1 Graphical representation – variables and intended interpretations

For the purpose of developing our Bayes net model of pharmacological inference, our main interest is determining rational degrees of belief in the causal hypothesis "Drug D *causes* harm H in population U". For ease and clarity of exposition, we here use a binary propositional variable. We hence introduce a variable © with the intended meaning of © $= TRUE$ is that: "Drug D causes harm H in U".

As outlined in Section 3, we take it that rational belief in the causal hypothesis is based on causal indicators: difference-making, probabilistic dependence, dose-response relationship, mechanisms etc. We use binary propositional indicator variables IND_i ($i \in \{1, 2, 3, 4, ...\}$) with $IND_1 = \Delta, IND_2 = \text{PD}, IND_3 = DR, IND_4 = RoG, IND_5 = \text{M}, IND_6 = \text{T}$ (difference-making, probabilistic dependence, dose-response relationship, rate of growth, mechanisms and temporality, respectively) and possibly further variables for causal indication IND_k. Ind_i means that the i-th consequence of the causal hypothesis holds.

For every item of evidence r and every causal indicator i the item of evidence informs us about, we use a report variable REP_r^i.[19] Reports may come from various kinds of studies such as case reports,[20] case series, case-control studies and cohort

[19] Superscripts are suppressed in the notation, whenever no confusion arises.

[20] Case reports may contribute in two main different ways to harm assessment: 1) the first one(s) contribute to hypothesis generation: they function as alarm signals by identifying identify a previously unknown side effect; 2) following these hypothesis generation events, the subsequent case reports contribute to "strengthen" the signals, i.e., they have a confirmatory role, analogously to compared studies and other statistical evidence as illustrated in this paper. Other kinds of studies may also function as generators of hypotheses of course, but this role is mainly covered by case reports.

studies; or from experiments at various levels: in vitro, in vivo and clinical studies. Relevant evidence may also come from knowledge discovery techniques as well as computational modeling. The meaning of Rep_r^i is that the item of evidence is consistent with the causal indicator for a population for which the studied population is representative.

We use two further variables for relevance and reliability for report variables in order to account for these two dimensions of the evidence and their role in (dis)confirming the investigated hypothesis (cf. Section 3.3).

5.2 Probabilistic independencies

By borrowing the quite general reconstruction of scientific inference from Bovens & Hartmann we import their direct and indirect probabilistic dependencies and conditional independencies (see our summary of the relevant independencies from Bovens and Hartmann (2003) in Section 4.2e). Beyond that, our choice of causal indicators and subsequently their formalisation as variables (i.e., nodes in the graph) requires us to make the theoretical, implicatory dependencies transparent by expressing them as links in the network. The following list motivates our modeling choices for the formalisation of the conceptually related indicators RoG, PD, DR, and Δ:

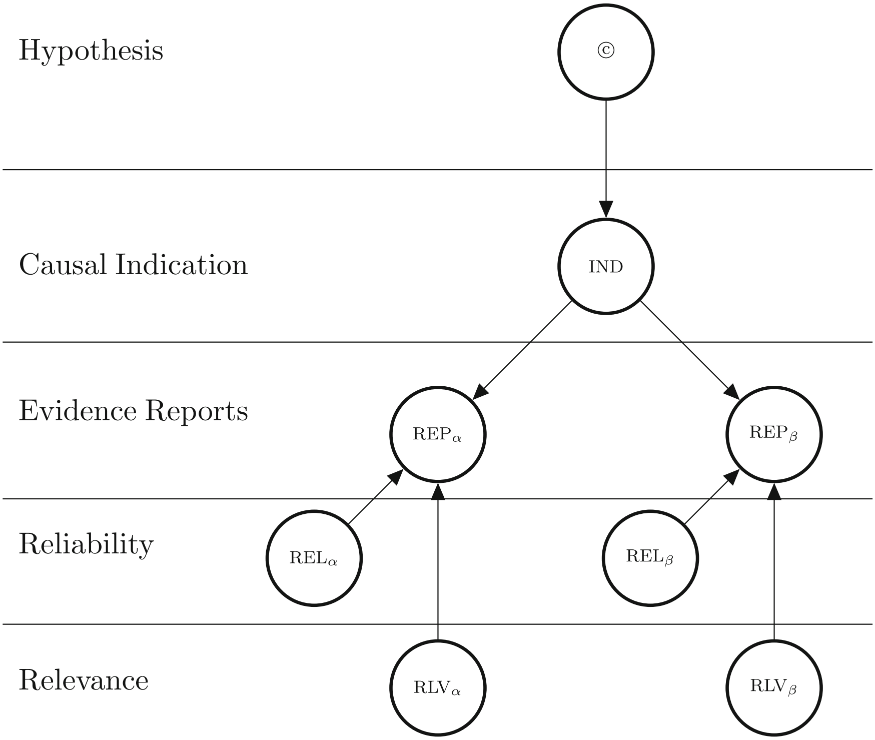


Fig. 8 Graph structure of the Bayesian network for two reports and epistemic categories

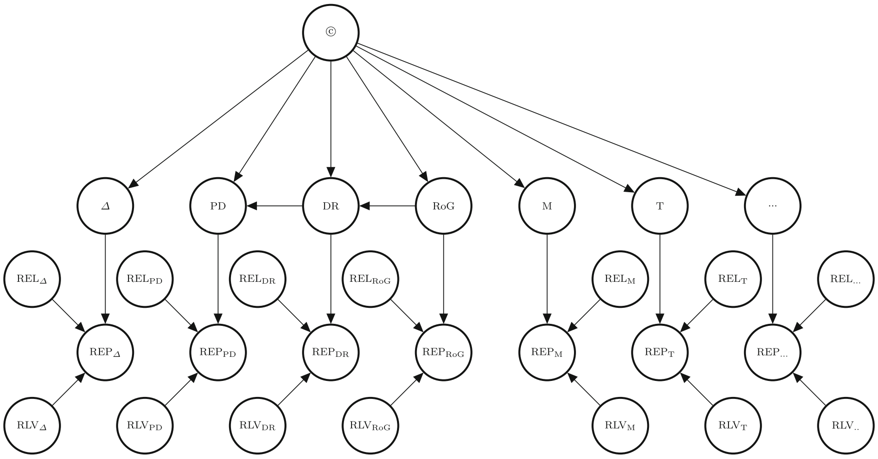


Fig. 9 Graph of the Bayesian network with one report for every causal indicator variable. The *dots* indicate that there might be further indicators of causality not considered here

1. We model the implication relations between the four indicators as edges on the second level of our network: the detection of a high rate of growth implies a dose-response relationship which in turn means that the variables under consideration are probabilistically dependent.

   Note that the edges on the second level are not a superfluous addition: If probabilistic dependence is measured but the causal hypothesis is known to be false (i.e., © is fixed to $FALSE$), the variables PD, DR, and RoG remain dependent since they overlap conceptually. Inserting direct edges on the second level precisely expresses this overlap. The *categorical independence* between the hypothesis level and the indicator level becomes apparent in this structure: PD, DR, and RoG are not dependent *via* the © node, but directly linked to one another on the indicator level.
2. No direct edge links RoG and PD since an observed high rate of growth implies an observed dose-response which implies an observed probabilistic dependence in turn and mediates the inference from RoG to PD (in other words, DR screens off PD from RoG).
3. Our choice to not insert an edge between DR, RoG, PD and Δ reflects our intention to clearly demarcate the conceptual/methodological dividing line between observational/static and interventional/dynamic support for the causal hypothesis: © screens off Δ from the observational/static indicators.[21] And formally: $\Delta \perp\!\!\!\perp DR, RoG, PD \mid$ ©. This principled distinction is already laid out in Hume's famous twofold definition of causation which can be seen as a

[21] By 'observational/static' we refer to inference from observation alone, whereas by 'interventional/dynamic' we refer to inference from data collected in interaction with the investigated system or population. For example, this contrast becomes evident in the difference between standard probabilistic conditioning (which amounts to shifting the focus in a probabilistic model) and conditioning with Pearl's *do*-operator (which amounts to transforming the probabilistic model).

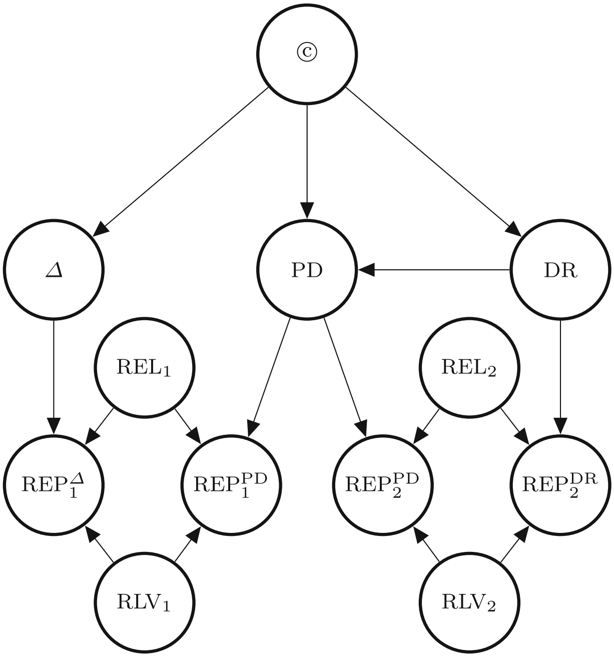


Fig. 10 Example of a Bayesian net for two studies which both inform us about two indicators. The study generating reports for difference making and probabilistic dependence may be an RCT, while the study generating the two reports on the right may be a case control study. For both studies, we use one reliability and one relevance variable each

point of reference both for regularity/supervenience as well as for counterfactual/manipulationist theories of causation (cf. Hume 1748, Sec. VII). Finally, all indicators listed above are imperfect ones, except for difference-making. Nevertheless, we are not collapsing Δ and © into a single node: Following the philosophical literature on causality, we consider that when a difference-making relationship between two events or variables holds, then this is a sufficient – although not necessary – condition for causality. This can be characterized in logical terms as an entailment relationship: $\Delta \supset$ ©. Hence, in our system, the probability of a causal relationship, given a genuine difference making relationship is 1: $P(© \mid \Delta) = 1$. The inverse entailment, © $\supset \Delta$, does not hold however. Although Δ – representing the possibility of *ideal controlled variance* – implies © in a definitional way, knowledge of © does not necessitate the existence difference-making – e.g., in cases of "holistic causation".[22]

[22] The latter case makes the conceptual divide even more obvious: If one knows the hypothesis to be true, learning that there is no difference-making would not change one's belief in a positive dose-response. In this case the causal relation under investigation would then be explained as holistic causation. We are thankful to an anonymous reviewer for pointing this case out to us.

Note also that we are purposely choosing to direct the edge between © and M towards M: We understand the existence of a mechanism as a testable consequence of the causal hypothesis, i.e., as a constitutive element of © rather than a pre-requisite (or even somehow causally prior) – in accordance with all other indicator nodes.[23]

The Figs. 8–10 graphically represent aspects of the graph of our Bayesian network. Figure 8 displays epistemic dimensions at stake, Fig. 9 shows the causal indicators, their reciprocal relations, and the studies which inform us about single indicators. Figure 10 depicts a case in which studies are informative about more than one indicator.

5.3 Adopting a prior

Success of Bayesian reasoning hinges on the choice of a suitable prior. Incorporating domain knowledge plays a major role in the choice of a prior. Domain knowledge may be elicited from experts.[24]

While the prior has to satisfy the conditional independences discussed above and incorporate prior domain knowledge, there are further properties in the problem specification that a sensible prior has to satisfy; which we shall now discuss.

$$\begin{aligned}
P(©) &\leq P(©|Ind) && \text{for all indicator variables} && (9)\\
P(©) &> P(©|\neg Ind) && \text{for all indicator variables} && (10)\\
P(©|Ind_i \& Ind_k) &\geq P(©|Ind_i) && \text{for all } i \neq k && (11)\\
P(©|Ind_i \& \neg Ind_k) &\leq P(©|Ind_i) && \text{for all } i \neq k && (12)\\
P(©|\Delta),\ P(©|DR) &> P(©|PD) && && (13)\\
P(©|\neg PD) &\leq P(©|\neg DR),\ P(©|\neg RoG) && && (14)\\
P(Ind_i|Rep_i \& Rlv_i \& Rel_i) &> P(Ind_i) && \text{for all } i && (15)\\
P(\neg Rep|\neg(Rel \& Rlv) \& Ind) &> P(\neg Rep|Ind). && && (16)
\end{aligned}$$

In Eq. 16 all report, reliability and relevance variables pertain to the same indicator variable.

Equation 9 means that the conditioning on one causal indicator boosts the belief in the causal hypothesis being true, while Eq. 10 means that conditioning on the negation of an indicator lowers the belief in the causal hypothesis. Similarly, the same holds in the presence of another instantiated indicator, see Eqs. 11 and 12. Equations 13 and 14 express that probabilistic dependence is a weaker causal indicator than *difference-making* or *high rate of growth*. Consequently, conditioning on

[23] We are thankful to an anonymous reviewer for hinting at sources of potential disagreement about the role of mechanistic knowledge and thus helping us elucidate our point here. For a discussion of different network structures and their use for hypothesis confirmation see, e.g., Wheeler and Scheines (2013).

[24] Elicitations of parts of priors from experts in a medical context has recently been reviewed in Johnson et al. (2010). Determination of prior distributions combining expert opinion with historical data is reported in Hampson et al. (2014, Section 4).

difference-making, dose-response relationship, or high rate of growth gives a greater boost to the belief in the causal hypothesis than probabilistic dependence. Vice versa, conditioning on the negation of probabilistic dependence reduces belief in the causal hypothesis more strongly than conditioning on the negation of a dose-response relationship or high rate of growth.

One reliable study, which is reliable and relevant for the target population, and finds that, say, there is probabilistic dependence between the drug an adverse drug reactions, significantly boosts the belief that there is probabilistic dependence in the target population; (15).

Equation 16 formalises the following thought: Belief in the inconsistency of a report and the respective indicator is boosted if the study is either irrelevant, unreliable, or both ($\neg(Rel\&Rlv)$ serves to explain the inconsistency).

5.4 Two model calculations

We now illustrate how to use our model by giving two example calculations. First, we show how to decide between withdrawing the drug or not. Secondly, we show how the posterior weight of an item of evidence that conflicts with a large consistent body of evidence decreases.

5.4.1 The threshold p^*

With the prior in place, we can now investigate conditions under which we will recommend to withdraw the drug D by calculating the threshold p^*.

We find for a pair of distinct indicator variables $IND_j, IND_k \neq \Delta$ which are not linked by an arrow

$$\begin{aligned} P(©|Ind_j\&\neg Ind_k) &= \frac{P(©\&Ind_j\&\neg Ind_k)}{P(Ind_j\&\neg Ind_k)} \\ &= \frac{P(©\&Ind_j\&\neg Ind_k)}{P(Ind_j\&\neg Ind_k\&©) + P(Ind_j\&\neg Ind_k\&\neg©)} \\ &= \frac{P(©\&Ind_j\&\neg Ind_k)}{P(Ind_j\&\neg Ind_k\&©) + P(Ind_j\&\neg Ind_k\&\neg©)}. \end{aligned}$$

Letting $a, b \in [0, 1]$

$$\begin{aligned} a &:= P(©\&Ind_j\&\neg Ind_k) = P(Ind_j|©) \cdot P(\neg Ind_k|©) \cdot P(©) \\ b &:= P(\neg©\&Ind_j\&\neg Ind_k) = P(Ind_j|\neg©) \cdot P(\neg Ind_k|\neg©) \cdot P(\neg©) \end{aligned}$$

we obtain

$$P(©|Ind_j\&\neg Ind_k) = \frac{a}{a+b}.$$

For fixed a, $\frac{a}{a+b}$ is a strictly decreasing function in b; if $a = 0$, then $\frac{a}{a+b} = 0$ (as long as $b \neq 0$ in which case the fraction is ill-defined), if $a = 1$, then $\frac{a}{a+b}$ varies between 1 and 0.5. For fixed $b > 0$, $\frac{a}{a+b}$ increases strictly with increasing a. If $b = 0$, then $\frac{a}{a+b} = 1$ (assuming $a \neq 0$), if $b = 1$, then $\frac{a}{a+b}$ varies between 0 and 0.5.

For the threshold p^* given by the decision problem ($p^* = P(©|Ind_j\&\neg Ind_k)$), we can compute the threshold value a^* for for fixed b (and vice versa) as follows (assuming we never divide by zero):

$$a^* = b \cdot \frac{p^*}{1-p^*} \qquad b^* = a \cdot \frac{1-p^*}{p^*} \quad .$$

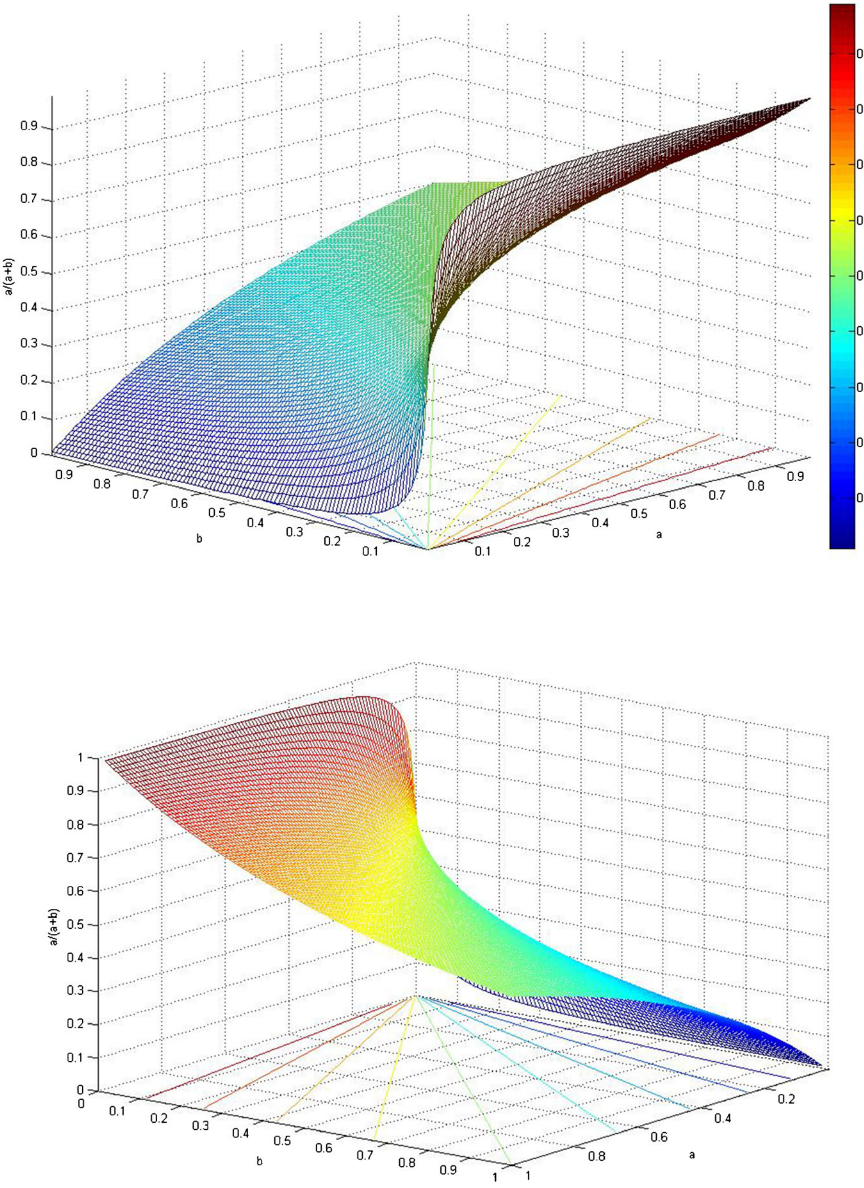


Fig. 11 Plot of probability $P(©|Ind_j\&\neg Ind_k)$ in dependence of a and b. *Contour lines* are displayed in the $a-b$-plane

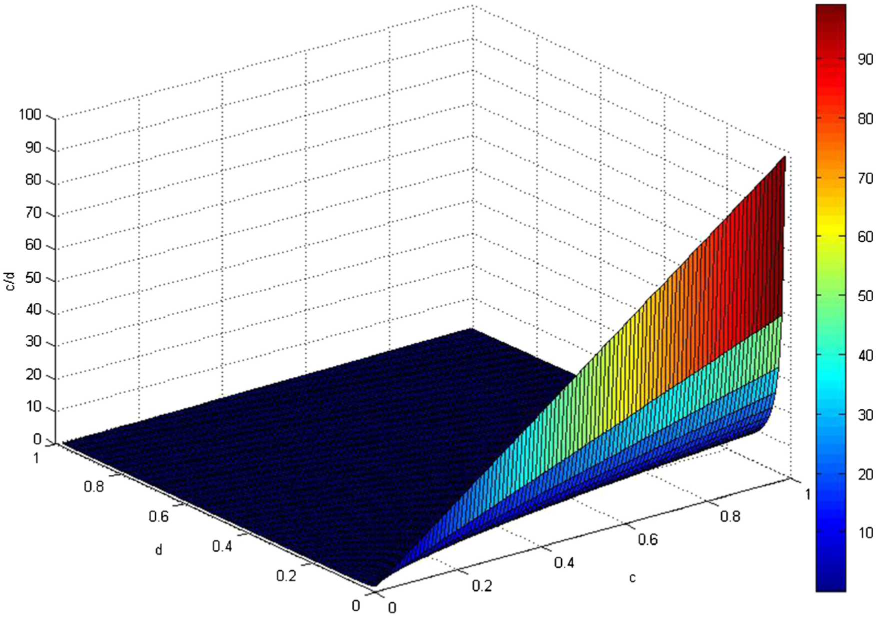


Fig. 12 Posterior probability of $\neg\varrho$ in dependence of c, d

For fixed b, if $a > a^*$, then our belief in the causal hypothesis, $D©H$, is too great to recommend the drug for use. For fixed a, if $b > b^*$, then our belief in the causal hypothesis, $D©H$, is low enough to recommend the drug for use.

The posterior probability $P(©|Ind_j\&\neg Ind_k)$ and the contour lines are plotted in Fig. 11. The contour lines $a^* = b \cdot \frac{p^*}{1-p^*}$ and $b^* = a \cdot \frac{1-p^*}{p^*}$ are linear curves which would intersect at the origin of the $a-b$-plane, if they were defined there. The clustering of the contour lines near the origin indicates that small changes in a or b, if a and b are small, can have a large effect on $\frac{a}{a+b}$, on the probability of $D©H$ and hence easily change the recommended action.

5.4.2 Dynamics

Conflicting reports are a fact of life in pharmacological inference. We here show that a report which conflicts with all other reports – which speak to the same causal indicator – has high posterior probability that it is not reliable or not relevant or both, given that the conflicting report comes from an independent source.

Let REP_j for $j \in \{0, 1, \ldots, N\}$ all concern indicator IND_k be such that report zero says that the Study 0 is *inconsistent* with the causal indicator IND_k while all other reports say that the observations are consistent with the causal indicator IND_k. Denote the proposition which captures the evidence other than the evidence obtained by Study 0 by

$$\rho := \bigwedge_{j=1}^{N} Rep_j\&Rel_j\&Rlv_j.$$

Denoting by $\varrho := Rel_0\&Rlv_0$ we now compute the posterior probability of $\neg\varrho$, that is, the posterior probability that either Study 0 is unreliable, irrelevant or both. In mathematical prose, the above-mentioned independence is expressed as $P(\rho\&\neg\varrho\&\neg Rep_0|Ind_k) = P(\rho|Ind_k) \cdot P(\neg\varrho\&\neg Rep_0|Ind_k)$.

Since $P(\rho\&\neg Ind_k)$ is *very* close to zero we find

$$\begin{aligned}
P_{posterior}(\neg\varrho) &= P(\neg\varrho|\neg Rep_0\&\rho) \\
&= \frac{P(\neg\varrho\&\neg Rep_0\&\rho\&Ind_k) + P(\neg\varrho\&\neg Rep_0\&\rho\&\neg Ind_k)}{P(\neg Rep_0\&\rho)} \\
&\geq \frac{P(\neg\varrho\&\neg Rep_0\&\rho\&Ind_k)}{P(\neg Rep_0\&\rho)} \\
&= \frac{P(Ind_k)P(\neg\varrho\&\neg Rep_0|Ind_k)P(\rho|Ind_k)}{P(\neg Rep_0\&\rho\&Ind_k) + P(\neg Rep_0\&\rho\&\neg Ind_k)} \\
&= \frac{P(Ind_k)P(\neg\varrho\&\neg Rep_0|Ind_k)P(\rho|Ind_k)}{P(Ind_k)P(\neg Rep_0|Ind_k)P(\rho|Ind_k) + P(\neg Ind_k)P(\neg Rep_0|\neg Ind_k)P(\rho|\neg Ind_k)} \\
&\approx \frac{P(Ind_k)P(\neg\varrho\&\neg Rep_0|Ind_k)P(\rho|Ind_k)}{P(Ind_k)P(\neg Rep_0|Ind_k)P(\rho|Ind_k)} \\
&= \frac{P(\neg\varrho\&\neg Rep_0|Ind_k)}{P(\neg Rep_0|Ind_k)} \\
&= \frac{P(\neg\varrho\&\neg Rep_0\&Ind_k)}{P(\neg Rep_0\&Ind_k)} \\
&= P(\neg\varrho|\neg Rep_0\&Ind_k).
\end{aligned}$$

The posterior probability of ϱ being false is boosted the larger $c = P(\neg\varrho\&\neg Rep_0\&Ind_k)$ is compared to $d = P(\neg Rep_0\&Ind_k)$ as depicted in Fig. 12. For a body of evidence consistent with, say, probabilistic dependence, and a report which does not report a probabilistic dependence between drug and harms it is a priori (highly) likely, that this report is either not reliable or not relevant or both.

6 Conclusions

6.1 Summary

The here proposed Bayesian network allows the amalgamation of various pieces of evidence from heterogeneous sources and methods and providing an overall estimate of the causal hypothesis. In particular, our approach

i) identifies possible indicators of causality on the basis of the methodological and philosophical literature on causality, evidence, and causal inference;
ii) embeds them in a topological framework of probabilistic dependencies and independencies grounded in assumptions regarding their reciprocal epistemic interconnections;
iii) weakly orders some of these probabilistic dependencies as a function of their inferential strength with respect to the confirmation of causal hypotheses.

For this, we have adopted Bovens and Hartmann (2003) proposal to use Bayesian confirmation theory in order to account for (and mathematically explain) some phenomena related to scientific inference; such as the confirmatory power of the coherence of the body of evidence, the epistemic interaction of consistency of measurements and reliability of information sources, as well as the modular contribution of different "lines of evidence" related to diverse observable consequences of the investigated hypothesis. We have adapted this framework to situations of causal inference and consequently specified a concrete structure for that purpose. We have then illustrated its epistemic and heuristic virtues as an instrument for evidence amalgamation, in the context of causal inference of drug-induced harm.

Our approach thereby satisfies the desiderata listed at the end of Section 2.2: probabilistic hypothesis confirmation, incorporation of heterogeneous kinds of data, facilitation of diverse types of inferential patterns (more on this on future work below) with a particular focus on causal assessment in pharmacology.

6.2 Discussion

6.2.1 Limitations

Our model, as well as every other model, is a simplification of the phenomenon of interest which entails a regrettable but unavoidable loss of information. One simplification was the use of binary propositional variables which stands in tension with concepts such as probabilistic dependence and rate of growth which clearly come in degrees. For the sake of simple exposition and tractable calculation, we made this simplification. Nothing hinges on this, the machinery of Bayesian networks can be applied to non-binary variables in the same manner. In general, the more information available the stronger the need for variables with more values.

Furthermore, we have limited the model to indicators of causality in pharmacology which derive from Hill's guidelines. We freely admitted in Section 3.2 that there may be further indicators. No suggestion is made here that all these further indicators can be accounted for in the model; in the outlook below we talk about a relevant inference which cannot be drawn within our model.

6.2.2 Virtues and context

Pragmatically, the here presented model has the virtue of being computationally simple in the following sense. Defining prior probabilities on rich structures can be a hard practical problem. A graphical representation of the conditional independencies in terms of the graph of Bayesian network allows one to specify the full prior by specifying all conditional probabilities at the variables. For a relatively sparse graph such as ours (the number of parents of every variable is at most three), specifying all conditional probabilities requires far less input than specifying the full prior by assigning probabilities to all states.

With respect to other proposals for evidence evaluation and causal assessment, our approach has the following virtues:

1. Our method accommodates many – sometimes only apparently contrasting – intuitions already expressed by philosophers of medicine regarding standard approaches to evidence evaluation.
2. Our method allows for many inferential patterns to contribute to the overall causal assessment of drug-induced harm (coherence, consistency, reasoning by analogy, etc.) and explicitly (as well as formally) accommodates these patterns in the belief propagation network.

We think that our framework can help with the first point for the following reasons: 1) by breaking down the evidential line between pieces of evidence and causality into a two-stage process mediated by causal indicators, we help disentangle philosophical issues related to the conceptualisation of causality from those related to causal inference and diagnostics. We believe that some disputes among philosophers and methodologists may be solved by keeping these levels as distinct; 2) our framework aims to probabilistic causal assessment, hence bypasses problems generated by accounts that aim to establish causal claims categorically; 3) we reserve separate nodes for relevance and reliability issues, which also generate much discussion in the methodological literature, and are at the root of concerns about the value of one kind of evidence rather than the other.

Most of the debate revolves around the prominence of randomized clinical trials in Evidence Based Medicine vs. the opposing view that other forms of evidence are also necessary/important for the purpose of causal assessment, for various reasons. Speaking in our terms, the implicit assumptions underpinning the EBM viewpoint is that ideal RCTs (that is internally valid ones) are perfect indicators of difference making, and difference making is a perfect indicator of causality. This can be represented in logical terms as an entailment relationship: $\mathrm{RCT} \supset \Delta \supset \copyright$. Various objections have been raised against these assumptions from the pluralists side, however the debate conflates the two inclusion relationships into one: that is, what is discussed is whether RCTs provide perfect information for causality (directly): $\mathrm{RCT} \supset \copyright$.

We present here how our framework can accommodate intuitions coming from both sides of the dispute, by considering five specific issues: the EBM vs. pluralist approach to causal inference, evidence hierarchies, causal holism, relevance (external validity), and reliability. These issues also show how our framework provides a higher order perspective on these debates by effectively embedding these various epistemic dimensions in a concrete topology.

6.2.3 Causal indicators: EBM vs. pluralists debate

By privileging the role of RCTs for the purpose of establishing causal claims, the EBM paradigm implicitly takes difference making as a perfect indicator for causality ($\Delta \supset \copyright$)[25], and the other indicators as very weak ones; hence it concentrates its efforts on having as reliable as possible evidence for that kind of indicator.

[25] This directly derives from the potential outcome approach underpinning RCT methodology. See Holland (1986), Rubin (2011), and Vandenbroucke et al. (2016) for a critical appraisal of this approach.

The contending view is that different indicators may have complementary epistemic roles in supporting the hypothesis of causality. This is held for instance by the "new mechanists" (henceforth the "Kentians"), led (among others) by Jon Williamson. For instance, Clarke et al. (2014) claim that evidence about difference making helps in de-masking causes which might be canceled out by back-up/compensatory mechanisms in the organ system, whereas evidence about mechanisms is needed in order to design and interpret statistical studies. Hence, such different kinds of evidence reciprocally support each other and jointly (dis-)confirm the causal claim under investigation. This proposal stems from Russo and Williamson (2007), where both statistical evidence at the phenotypic/population level and evidence about (molecular-cellular) mechanisms is required to establish causal claims (hence, they are jointly necessary and sufficient to establish causality).

Howick expresses the opposing view that: "There are many cases where patient-relevant effects of medical therapies have been established by comparative clinical studies *alone*." (Howick 2011, p. 939).

In our terms we can take Howick to hold the view that $P(\copyright \mid \Delta)$ is 1, and that ideal RCTs provide strong evidence for Δ, $P(\Delta \mid RCT) \approx 1$ hence offering strong enough evidence to establish the causal claim, while for the Kentians $P(\copyright \mid \Delta)$ is too small to establish the causal claim. For them, also having mechanistic evidence is required to establish the causal claim: $P(\copyright \mid \Delta \& M) \approx 1$.

Our framework allows for RCT $\supset \Delta \supset \copyright$ to hold, but the entailment relationship between RCTs and Δ is one between ideal RCTs and Δ. Hence in any concrete case $P(\Delta|\text{RCT}) < 1$, and this leaves room for other kinds of evidence to also contribute to hypothesis confirmation. On the other side, our approach relaxes the Kentian theory by dropping any necessity or sufficiency requirements for causal inference. This approach is therefore much more flexible and responds both to the EBM intuition underpinning the privileged role assigned to RCTs, as well as to the pluralist intuition that various kinds of evidence are contributory to causal assessment.

6.2.4 Evidence hierarchies

In relation to evidence hierarchies – which is a strong point of contention among philosophers of medicine and methodologists – our inequalities (13), (14), nicely parallel the ranking proposed there.

Evidence hierarchies have been developed as a decision tool to help clinicians pressed by time constraints, to integrate their clinical expertise with evidence coming from basic and clinical research (Guyatt et al. 1992; Sackett et al. 1996; Straus and McAlister 2000). In these rankings, randomised studies are *ceteris paribus* preferred to non-randomised studies.[26] However, also the strength of the effect magnitude

[26] The rationale for this ranking is provided by methodological-foundational considerations mainly developed within standard statistics and follow a kind of hypothetico-deductive approach to scientific inference (see also our comments on Experiment above, Page 19).

and dose-response gradient are considered essential features in evaluating evidence (Howick 2011; Glasziou et al. 2007). Hence, our inequality constraints mirror the categorical ordering recommended in such hierarchies.

What differentiates our framework from standard evidence rankings however is that these have predominantly been formalised as lexicographic decision rules. This means that higher-level studies trump lower-level ones: when two studies of different levels deliver contradictory findings, then the one higher in the evidence hierarchy is considered more reliable and one is allowed to discard the lower level one.[27] Our framework incapsulates the rationale for ranking evidence in Eqs. 13 and 14 but at the same time allows one to take into account all evidence, and to act accordingly, as soon as the probability of the causal hypothesis goes above the threshold established by the other dimensions of the decision (utility of withdrawing/not withdrawing the drug, conditional on the probability of it causing the suspected harm).[28]

6.2.5 Causal interaction and causal holism vs. modularity

Methodological pluralists such as Cartwright (Cartwright and Stegenga 2011; Cartwright 2007a), and Stegenga (2014), among others, express concerns against the privileged role of RCTs also on grounds that classical 'linear' approaches to causal inference cannot do justice to the complexity of causal phenomena in the biological and social sciences, characterized by nonlinear causation and causal interactions. In the same line, also modular conceptualization of causes such as the ones implied in the causal graph methodology developed by Pearl (2000) and Glymour (Spirtes et al. 2000) and colleagues (see also Woodward 2003), are under attack for failing to recognize that causes may be holistic and therefore may be not adequately captured by a difference making account.

Strictly speaking this sort of criticism does not deny that $\Delta \supset ©$, but only denies the reverse: $© \supset \Delta$. However, in the causal graph literature the defining features for causality jointly entail that $\Delta \Leftrightarrow ©$. We contribute to the debate by not collapsing Δ and © into a single node.

We consider that when a difference-making relationship between two events or variables holds, then this is a sufficient – although not necessary – condition for causality. This can be characterized in logical terms as an entailment relationship: $\Delta \supset ©$. Hence, in our system, the probability of a causal relationship, given a genuine difference making relationship is 1: $P(© \mid \Delta) = 1$. The inverse entailment though, $© \supset \Delta$, does not hold: although Δ – representing the possibility of *ideal controlled*

[27] A somewhat unwanted consequence of this "take the best" approach is that it has become commonplace to assume an uncommitted attitude towards observed associations least they are "proved" by gold standard evidence (see the still ongoing debate on the possible causal association between paracetamol and asthma; (Shaheen et al. 2000; Eneli et al. 2005; Shaheen et al. 2008; Henderson and Shaheen 2013; Allmers et al. 2009; McBride 2011; Heintze and Petersen 2013; Martinez-Gimeno and García-Marcos 2013)).

[28] This also complies with the precautionary principle in risk assessment and with how decisions should be made in health settings (see Section 2.2).

variance – implies © in a definitional way, knowledge of © does not necessitate the existence difference-making – e.g., in cases of “holistic causation”.[29]

6.2.6 Relevance

The ‘Kentians’ also maintain that a sample which is small compared to the target population on its own is *not* sufficient to license causal claims about the target population (problem of external validity). Hence, evidence that there exists some mechanism responsible for the phenomenon in the observed sample which is also present in the target population is required, cf. Russo and Williamson (2007), This is also a point stressed by Cartwright (Cartwright and Stegenga 2011).[30] We formally distinguish the role of evidence of mechanisms for the purpose of causal assessment from its role for the purpose of establishing external validity by associating the latter to our relevance node RLV. This allows to explicitly distinguish the different kinds of inductive risk involved in the inference: 1) an inductive leap from causal indicator (be it correlation or mechanistic evidence) to causality, 2) another one from inferring causality in a given population/model to establishing it in another population/model (see also Poellinger 2017). More importantly, adding a relevance node to the evidence reports allows for even highly reliable RCTs to provide little evidential support if they are not considered to be relevant.

6.2.7 Reliability

A pragmatic defence of RCTs, which nevertheless acknowledges their methodological limitations can be found in La Caze et al. (2012), La Caze (2009), and Teira (2011). While La Caze provides reasons to privilege RCTs over non-randomised studies in terms of reliability and computational tractability, Teira analyses the issue within the perspective of strategic behaviour and regulatory constraints (see also Teira and Reiss 2013). Teira acknowledges the methodological limitations attributed to RCTs by philosophers of science, however, he goes on, randomisation ”is still a warrant that the allocation was not done on purpose with a view to promoting somebody’s interests”. Thus, he explains the success of RCTs in regulatory protocols for market approval of pharmaceutical products on grounds that they guarantee impartiality. Randomization serves the purpose to avoid that the uncertainty related to causal inference be advantageously exploited by one party or the other. By introducing a reliability node REL, and thereby breaking up the different dimensions of evidence (strength, relevance, reliability) we allow for them to be explicitly tracked

[29] This also responds to concerns expressed in Cartwright (2007b), Mumford S. and Anjum (2011), Anjum and Mumford (2012), and Kerry et al. (2012).

[30] Both the Kentians and Cartwright construe the term “mechanism” in slightly different fashion than we did here. For them, a mechanism need not be described on the (sub-)molecular level. This detail is not relevant for our current discussion.

in the body of evidence. This makes it possible to parcel out the strength of evidence from the method with which it was obtained.[31]

6.2.8 Higher order perspective on the methodological debate

Furthermore, the Bayesian network and its nodes – representing epistemic categories (relevance, reliability, various causal indicators, etc.) – provide us with a greater insight into the philosophical dissent around EBM. For instance whereas Worral's criticism, see Worrall (2007a, 2010) against the privilege accorded to RCTs for causal inference in EBM insists on questioning their high reliability, Cartwright's view (Cartwright and Stegenga 2011) is that they provide very limited information on the effect of the intervention in other populations than the study population ("external validity"). Hence, these criticisms address different nodes in the causal inference, although they regard the same study type. With this, our framework provides a higher order perspective on these debates by effectively embedding these various epistemic dimensions in a concrete topology.[32]

This leads us directly to our second point: our approach allows for the integration of various kinds of evidence through various types of inferential strategies, in accordance with more general concerns expressed by pluralists. Stegenga, for instance, explicitly mentions Hill's viewpoint on causal inference and claims that: "a plurality of *reasoning strategies* appealed to by the epidemiologist Sir Bradford Hill is a superior strategy for assessing a large volume and *diversity* of evidence" (Stegenga 2011) (emphasis added). Indeed, the framework presented here also provides a fruitful platform for integrating insights developed in the philosophy of science around such topics as the role of replication in assessing the reliability of evidence (Open Science Collaboration 2015; Meehl 1990; Lamal 1990; Hempel 1968; Platt 1964), as well as the confirmatory role of explanatory power (McGrew 2003; Crupi et al. 2013; Cohen 2016; Lipton 2003) and coherence (Dietrich and Moretti 2005; Moretti 2007; Wheeler and Scheines 2013; Fitelson 2003; Bovens and Hartmann 2003).

6.3 Outlook

We are engaged in two avenues of further research: Firstly, we are interested in validating our framework through applications through three case studies and computer simulations. Likely case studies are: i) the ongoing debate on the possible causal association between Paracetamol and asthma; ii) integration of mechanistic and statistical evidence in the Torcetrapib case; iii) the role of causal interaction and safety in the Terfenadine case.

[31] Osimani and Landes (Forthcoming) investigates the various concepts of reliability involved in such considerations.

[32] Moreover, our approach addresses explicitly the issue of external validity by formally incorporating reasoning by analogy (see Section 3.2.6).

This paper focuses on inference *within one model*, rooting in *one hypothesis*, but our framework allows for going beyond the network's 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. To give one prominent example: in the current debate, the *No Alternative Argument* (NAA) is discussed as lending valid confirmatory support to a specific hypothesis through the absence of a better candidate (also see the comments on safety assessment by the FDA in Food Drug Administration (2009, p. 5). Our initial decision to employ the Bayes net framework in the pharmacological context comes to fruition once more when we insert networks like Fig. 9 in highly expressive NAA metastructures (as suggested, e.g., by Dawid, Hartmann, and Sprenger in Dawid et al. 2015) for justifying a working hypothesis when no available evidence adjudicates conclusively. We leave this for another paper.

Acknowledgments This paper was presented at various workshops and conferences in Munich, New Brunswick, Sheffield, Helsinki, Durham, Amsterdam, and Ferrara. We greatly profited from the comments and suggestions made by the audiences; in particular we wish to thank Rani Lill Anjum, Timo Bolt, Giovanni Boniolo, Branden Fitelson, Bennett Holman, Phyllis Illari, Mike Kelly, Ulrich Mansmann, Carlo Martini, Julian Reiss, Stephen Senn, Beth Shaw, Jacob Stegenga, and Veronica Vieland. We also thank the focus group members of the ERC project "Philosophy of Pharmacology: Safety, Statistical Standards, and Evidence Amalgamation", to whom we owe a considerable improvement of the paper's argumentation: Jeffrey Aronson, Lorenzo Casini, Brendan Clarke, Vincenzo Crupi, Sebastian Lutz, Federica Russo, Glenn Shafer, Jan Sprenger, David Teira, and Jon Williamson. We are extremely grateful to our colleagues at the Munich Center for Mathematical Philosophy, who helped us clarify the objectives and scope of our research, and suggested possible paths of development; in particular we wish to thank Seamus Bradley, Richard Dawid, Samuel C. Fletcher, Stephan Hartmann, Alexander Reutlinger, and Gregory Wheeler. Finally, we thank two anonymous reviewers for their comments. These significantly helped us refine some important assumptions in our theoretical framework. Of course any inaccuracies or errors in the text are, however, our own responsibility.

References

Abernethy, D., & Bai, G. (2013). Systems pharmacology to predict drug toxicity: integration across levels of biological organization. *Annual Review of Pharmacology and Toxicology*, *53*, 451–73. doi:10.1146/annurev-pharmtox-011112-140248.

Allmers, H., Skudlik, C., & John, S.M. (2009). Acetaminophen use: a risk for asthma? *Current Allergy and Asthma Reports*, *9*(2), 164–7. doi:10.1007/s11882-009-0024-3.

Anjum, R.L., & Mumford, S. (2012). Causal dispositionalism. Properties, Powers and Structure 101–118, 7 In Bird, A., Ellis, B., & Sankey, H. (Eds.), Routledge.

Baetu, T.M. (2016). The 'Big picture': the problem of extrapolation in basic research. *British Journal for the Philosophy of Science*, *67*(4), 941–964. doi:10.1093/bjps/axv018.

Bartha, P. (2013). Analogy and analogical reasoning. In Zalta, E.N. (Ed.) The Stanford encyclopedia of philosophy, fall 2013 edn.

Bartha, P.F.A. (2010). By parallel reasoning: the construction and evaluation of analogical arguments. Oxford University Press.

Bes-Rastrollo, M., Schulze, M.B., Ruiz-Canela, M., & Martinez-Gonzalez, M.A. (2013). Financial conflicts of interest and reporting bias regarding the association between sugar-sweetened beverages and weight gain: a systematic review of systematic reviews. *PLOS Medicine*, *10*(12), 1–9. doi:10.1371/journal.pmed.1001578.

BonJour, L. (2010). *Epistemology. Classic problems and contemporary responses*. Rowman & Littlefield Publishers.

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford University Press.

Britton, O.J., Bueno-Orovio, A., Van Ammel, K., Lu, H.R., Towart, R., Gallacher, D.J., & Rodriguez, B. (2013). Experimentally calibrated population of models predicts and explains intersubject variability in cardiac cellular electrophysiology. *Proceedings of the National Academy of Sciences*, *110*(23), E2098–E2105. doi:10.1073/pnas.1304382110.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., & Munafo, M.R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376. doi:10.1038/nrn3475.

Carnap, R. (1947). On the application of inductive logic. *Philosophy and Phenomenological Research*, *8*(1), 133–148. http://www.jstor.org/stable/2102920.

Carné, X., & Cruz, N. (2005). Ten lessons to be learned from the withdrawal of Vioxx. *European Journal of Epidemiology*, *20*(2), 127–129. doi:10.1007/s10654-004-6856-1.

Cartwright, N. (2007a). Are RCTs the Gold Standard? *Biosocieties*, *2*, 11–20. doi:10.1017/S1745855207005029.

Cartwright, N. (2007b). Causal powers: what are they? Why do we need them? What can be done with them and what cannot? Tech. Rep 04/07. http://www.lse.ac.uk/CPNSS/research/concludedResearchProjects/ContingencyDissentInScience/DP/CausalPowersMonographCartwrightPrint.

Cartwright, N. (2008). Evidence-based policy: what's to be done about relevance? *Philosophical Studies*, *143*(1), 127–136. doi:10.1007/s11098-008-9311-4.

Cartwright, N., & Stegenga, J. (2011). A theory of evidence for Evidence-Based policy. In Dawid, P., & Twinning William Vasilaki, M. (Eds.) Evidence, Inference and Enquiry, chap. 11, OUP (pp. 291–322).

Chan, A.W., & Altman, D.G. (2005). Epidemiology and reporting of randomised trials published in PubMed journals. *The Lancet*, *365*(9465), 1159–1162. doi:10.1016/S0140-6736(05)71879-1.

Clarke, B., Leuridan, B., & Williamson, J. (2014). Modelling mechanisms with causal cycles. *Synthese*, *191*(8), 1651–1681. doi:10.1007/s11229-013-0360-7.

Cohen, M.P. (2016). On three measures of explanatory power with axiomatic representations. *British Journal for the Philosophy of Science*, *67*(4), 1077–1089. doi:10.1093/bjps/axv017. Early view.

Craver, C. (2007). *Explaining the brain: mechanisms and the mosaic unity of neuroscience*. Oxford: Clarendon Press.

Crupi, V., Chater, N., & Tentori, K. (2013). New axioms for probability and likelihood ratio measures. *British Journal for the Philosophy of Science*, *64*(1), 189–204. doi:10.1093/bjps/axs018.

Crupi, V.C., & Tentori, K. (2014). State of the field: measuring information and confirmation. *Studies in History and Philosophy of Science Part A*, *47*, 81–90. doi:10.1016/j.shpsa.2014.05.002.

Dardashti, R., Thébaut, K., & Winsberg, E. (2016). Confirmation via analogue simulation: what dumb holes can tell us about gravity. *British Journal for the Philosophy of Science*. doi:10.1093/bjps/axv010. Forthcoming.

Darden, L. (2006). *Reasoning in biological discoveries: essays on mechanisms, interfield relations, and anomaly resolution*. New York: Cambridge University Press.

Darwiche, A. (2009). *Modeling and reasoning with Bayesian networks*. Cambridge University Press.

Dawid, R., Hartmann, S., & Sprenger, J. (2015). The no alternatives argument. *British Journal for the Philosophy of Science*, *66*(1), 213–234. doi:10.1093/bjps/axt045.

Dietrich, F., & Moretti, L. (2005). On coherent sets and the transmission of confirmation. *Philosophy of Science*, *72*(3), 403–424. doi:10.1086/498471.

Doll, R., & Peto, R. (1980). Randomised controlled trials and retrospective controls. *British Medical Journal*, *280*, 44. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1600504/.

Dunn, A.G., Arachi, D., Hudgins, J., Tsafnat, G., Coiera, E., & Bourgeois, F.T. (2014). Financial conflicts of interest and conclusions about neuraminidase inhibitors for influenza. *Annals of Internal Medicine*, *161*(7), 513–518. doi:10.7326/M14-0933.

Edwards, W., Lindman, H., & Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242. doi:10.1037/h0044139.

Eneli, I., Katayoun, S., Camargo, C., & Barr, G.R. (2005). Acetaminophen and the risk of asthma. The epidemiologic and the pathophysiologic evidence. *CHEST*, *127*(2), 604–612. doi:10.1006/aama.1996.0501.

Fitelson, B. (2003). A probabilistic theory of coherence. *Analysis*, *63*(279), 194–199. doi:10.1111/1467-8284.00420.

Food Drug Administration (2009). Drug induced liver injury: premarketing clinical evaluation - guidance for industry. http://www.fda.gov/downloads/Drugs/Guidance/UCM174090.pdf.

Freedman, D. (1997). From association to causation via regression. *Advances in Applied Mathematics*, *18*(1), 59–110. doi:10.1006/aama.1996.0501.

Glasziou, P., Chalmers, I., Rawlins, M., & McCulloch, P. (2007). When are randomised trials unnecessary? Picking signal from noise. *British Medical Journal*, *7589*, 349–351. doi:10.1136/bmj.39070.527986.68.

Guyatt, G. et al. (1992). Evidence-based medicine: a new approach to teaching the practice of medicine. *Jama*, *268*(17), 2420–2425. doi:10.1001/jama.1992.03490170092032.

Hampson, L.V., Whitehead, J., Eleftheriou, D., & Brogan, P. (2014). Bayesian methods for the design and interpretation of clinical trials in very rare diseases. *Statistics in Medicine*, *33*(24), 4186–4201. doi:10.1002/sim.6225.

Heintze, K., & Petersen, K. (2013). The case of drug causation of childhood asthma: antibiotics and paracetamol. *European Journal of Clinical Pharmacology*, *69*(6), 1197–1209. http://dx.doi.org/10.1007/s00228-012-1463-7.

Hempel, C.G. (1968). Maximal specificity and lawlikeness in probabilistic explanation. *Philosophy of Science*, *35*(2), 116–133. http://www.journals.uchicago.edu/doi/abs/10.1086/288197.

Henderson, A.J., & Shaheen, S.O. (2013). Acetaminophen and asthma. *Paediatric Respiratory Review*, *14*(1), 9–15. doi:10.1016/j.prrv.2012.04.004.

Herxheimer, A. (2012). Pharmacovigilance on the turn? Adverse reactions methods in 2012. *British Journal of General Practice*, *62*(601), 400–401. doi:10.3399/bjgp12X653453.

Hesse, M.B. (1952). Operational definition and analogy in physical theories. *British Journal for the Philosophy of Science*, *2*(8), 281–294. http://www.jstor.org/stable/686017.

Hesse, M.B. (1959). On defining analogy. *Proceedings of the Aristotelian Society*, *60*, 79–100. http://www.jstor.org/stable/4544623.

Hesse, M.B. (1964). Analogy and confirmation theory. *Philosophy of Science*, *31*(4), 319–327. http://www.jstor.org/stable/186262.

Hill, A.B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, *58*(5), 295–300.

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945–960. doi:10.1080/01621459.1986.10478354.

Holman, B., & Bruner, J.P. (2015). The problem of intransigently biased agents. *Philosophy of Science*, *82*(5), 956–968. doi:10.1086/683344.

Horton, R. (2004). Vioxx, the implosion of Merck, and aftershocks at the FDA. *The Lancet*, *364*(9450), 1995–1996. doi:10.1016/S0140-6736(04)17523-5.

Howick, J. (2011). Exposing the vanities - and a qualified defense - of mechanistic reasoning in health care decision making. *Philosophy of Science*, *78*(5), 926–940. doi:10.1086/662561.

Howson, C., & Urbach, P. (2006). *Scientific Reasoning*, 3 edn. Open Court.

Hume, D. (1748). An enquiry concerning human understanding. The University of Adelaide Library 2004 (derived from the Harvard Classics Volume 37, 1910 P.F Collier & Son.) http://ebooks.adelaide.edu.au/h/hume/david/h92e/.

Ioannidis, J.P.A. (2016). Exposure-wide epidemiology: revisiting Bradford Hill. *Statistics in Medicine*, *35*(11), 1749–1762. doi:10.1002/sim.6825.

Johnson, S.R., Tomlinson, G.A., Hawker, G.A., Granton, J.T., & Feldman, B.M. (2010). Methods to elicit beliefs for Bayesian priors: a systematic review. *Journal of Clinical Epidemiology*, *63*(4), 355–369. doi:10.1016/j.jclinepi.2009.06.003.

Jüni, P., Nartey, L., Reichenbach, S., Sterchi, R., Dieppe, P.A., & Egger, M. (2004). Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. *The Lancet*, *364*(9450), 2021–2029. doi:10.1016/S0140-6736(04)17514-4.

Kerry, R., Eriksen, T.E., Lie, S.A.N., Mumford, S.D., & Anjum, R.L. (2012). Causation and evidence-based practice: an ontological review. *Journal of Evaluation in Clinical Practice*, *18*(5), 1006–1012. doi:10.1111/j.1365-2753.2012.01908.x.

Kment, B. (2010). Causation: determination and difference-making. *Noûs*, *44*(1), 80–111. doi:10.1111/j.1468-0068.2009.00732.x. Wiley Online Library.

Krumholz, H.M., Ross, J.S., Presler, A.H., & Egilman, D.S. (2007). What have we learnt from Vioxx? *British Medical Journal*, *334*(7585), 120–123. doi:10.1136b/mj.39024.487720.68.

Kuorikoski, J., Lehtinen, A., & Marchionni, C. (2010). Economic modelling as robustness analysis. *The British Journal for the Philosophy of Science*, *61*(3), 541–567. http://www.jstor.org/stable/40981302.

La Caze, A. (2009). Evidence-based medicine must be. *Journal of Medicine and Philosophy*, *34*(5), 509–527. doi:10.1093/jmp/jhp034.

La Caze, A., Djulbegovic, B., & Senn, S. (2012). What does randomisation achieve? *Evidence-Based Medicine*, *17*(1), 1–2. doi:10.1136/ebm.2011.100061.

LaFollette, H., & Shanks, N. (1995). Two models of models in biomedical research. *Philosophical Quarterly*, *45*(179), 141–160. http://www.jstor.org/stable/2220412.

Lamal, P. (1990). On the importance of replication. *Journal of Social Behavior and Personality*, *5*(4), 31–35.

Lewis, D. (1973). Causation. *Journal of Philosophy*, *70*(17), 556–567. http://www.jstor.org/stable/2025310.

Lewis, D. (1986). Causal explanation. In Philosophical papers, chap. 3. OUP, (Vol. II pp. 214–240).

Lewis, D. (2000). Causation as influence. *Journal of Philosophy*, *97*(4), 182–197. http://www.jstor.org/stable/2678389.

Lipton, P. (2003). *Inference to the best explanation*. Routledge.

Luján, J.L., Todt, O., & Bengoetxea, J.B. (2016). Mechanistic information as evidence in decision-oriented science. *Journal for General Philosophy of Science*, *47*(2), 293–306. doi:10.1007/s10838-015-9306-8.

Machamer, P., Darden, L., & Craver, C.F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25. http://www.jstor.org/stable/188611.

Martinez-Gimeno, A., & García-Marcos, L. (2013). The association between acetaminophen and asthma: should its pediatric use be banned? *Expert Review of Respiratory Medicine*, *7*(2), 113–122. doi:10.1586/ers.13.8.

McBride, J.T. (2011). The association of acetaminophen and asthma prevalence and severity. *Prediatrics*, *128*(6), 1–5. doi:10.1186/1745-6215-11-37.

McGrew, T. (2003). Confirmation, heuristics, and explanatory reasoning, *54*(4), 553–567. doi:10.1093/bjps/54.4.553.

Meehl, P.E. (1990). Appraising and amending theories: the strategy of lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141. doi:10.1207/s15327965pli0102_1.

Mill, J.S. (1884). *A system of logic, ratiocinative and inductive: being a connected view of the principles of evidence and the methods of scientific investigation*. Longmans, Green and Company.

Moretti, L. (2007). Ways in which coherence is confirmation conducive. *Synthese*, *157*(3), 309–319. doi:10.1007/s11229-006-9057-5.

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

Neapolitan, R.E. (2003). *Learning Bayesian networks*. Pearson.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *American Heart Journal*, *349*(6251), 943–aac4716–8. doi:10.1126/science.aac4716.

Osimani, B. (2007). *Probabilistic information and decision making in the health context: the package leaflet as basis for informed consent*. Doctoral Thesis, 1 edn Università della Svizzera Italiana.

Osimani, B. (2013). The precautionary principle in the pharmaceutical domain: a philosophical enquiry into probabilistic reasoning and risk aversion. *Health, Risk & Society*, *15*(2), 123–143. doi:10.1080/13698575.2013.771736.

Osimani, B. (2014a). Causing something to be one way rather than another. Genetic information, causal specificity and the relevance of linear order. *Kybernetes*, *43*(6), 865–881. doi:10.1108/K-07-2013-0149.

Osimani, B. (2014b). Hunting side effects and explaining them: should we reverse evidence hierarchies upside down? *Topoi*, *33*(2), 295–312. doi:10.1007/s11245-013-9194-7.

Osimani, B., & Landes, J. (Forthcoming). Exact replication or varied evidence? The varied of evidence thesis and its methodological implication in medical research.

Osimani, B., & Mignini, F. (2015). Causal assessment of pharmaceutical treatments: why standards of evidence should not be the same for benefits and harms? *Drug Safety*, *38*(1), 1–11. doi:10.1007/s40264-014-0249-5.

Osimani, B., Russo, F., & Williamson, J. (2011). Scientific evidence and the law: an objective bayesian formalisation of the precautionary principle in pharmaceutical regulation. *Journal of Philosophy, Science and Law*, 11. http://jpsl.org/files/9913/6816/1730/Bayesian-Formalization.pdf.

Papineau, D. (1993). The virtues of randomization. *British Journal for the Philosophy of Science*, *45*(2), 437–450. doi:10.1093/bjps/45.2.437.

Pearl, J. (2000). *Causality: models, reasoning, and inference*, 1st edn. Cambridge University Press.

Platt, J.R. (1964). Strong inference. *Science*, *146*(3642), 347–353. http://science.sciencemag.org/content/146/3642/347.

Poellinger, R. (2017). Analogy-based inference patterns in pharmacological research, Forthcoming.

Poellinger, R., & Beebe, C. (2017). Bayesian confirmation from analog models, Forthcoming.

Price, K.L., Amy Xia, H., Lakshminarayanan, M., Madigan, D., Manner, D., Scott, J., Stamey, J.D., & Thompson, L. (2014). Bayesian methods for design and analysis of safety trials. *Pharmaceutical Statistics*, *13*(1), 13–24. doi:10.1002/pst.1586.

Revicki, D.A., & Frank, L. (1999). Pharmacoeconomic evaluation in the real world. *PharmacoEconomics*, *15*(5), 423–434. doi:10.2165/00019053-199915050-00001.

Roush, S. (2005). *Tracking truth: knowledge, evidence, and science*. Oxford University Press.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. doi:10.1037/h0037350.

Rubin, D.B. (2011). Causal inference using potential outcomes. *Journal of the American Statistical Association*, *81*(396), 945–960. doi:10.1198/016214504000001880.

Russell, B. (1912). On the notion of cause. In Proceedings of the aristotelian society, (Vol. 13 pp. 1–26). http://www.jstor.org/stable/4543833.

Russo, F., & Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, *21*(2), 157–170. doi:10.1080/02698590701498084.

Sackett, D.L., Rosenberg, W.M., Gray, J.M., Haynes, R.B., & Richardson, W.S. (1996). Evidence based medicine: what it is and what it isn't. *Bmj*, *312*(7023), 71–72. doi:10.1136/bmj.312.7023.71.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.

Salmon, W. (1997). Causality and explanation: a reply to two critiques. *Philosophy of Science*, *64*(3), 461–477. http://www.jstor.org/stable/188320.

Schum, D. (2011). Classifying forms and combinations of evidence: Necessary in a science of evidence. In Dawid, P., Twinning, W., & Vasilaki, M. (Eds.) Evidence, inference and enquiry, chap. 2. OUP (pp. 11–36).

Senn, S. (2007). *Statistical Issues in Drug Development*. Wiley.

Shaheen, S., Potts, J., Gnatiuc, L., Makowska, J., Kowalski, M.L., Joos, G., van Zele, T., van Durme, Y., De Rudder, I., Wöhrl, S., Godnic-Cvar, J., Skadhauge, L., Thomsen, G., Zuberbier, T., Bergmann, K.C., Heinzerling, L., Gjomarkaj, M., Bruno, A., Pace, E., Bonini, S., Fokkens, W., Weersink, E.J.M., Loureiro, C., Todo-Bom, A., Villanueva, C.M., Sanjuas, C., Zock, J.P., Janson, C., & Burney, P. (2008). The relation between paracetamol use and asthma: a ga2len european case-control study. *European Respiratory Journal*, *32*(5), 1231–1236. doi:10.1183/09031936.00039208.

Shaheen, S., Sterne, J., Songhurst, C., & Burney, P. (2000). Frequent paracetamol use and asthma in adults. *Thorax*, *55*(4), 266–270. doi:10.1136/thorax.55.4.266.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search. Adaptive computation and machine learning*. MIT Press.

Steel, D. (2008). Across the boundaries. Extrapolation in biology and social sciences. Oxford University Press.

Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *42*(4), 497–507. doi:10.1016/j.shpsc.2011.07.003.

Stegenga, J. (2014). Down with the hierarchies. *Topoi*, *33*(2), 313–322. doi:10.1007/s11245-013-9189-4.

Stegenga, J. (2015). Measuring effectiveness. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *54*, 62–71. doi:10.1016/j.shpsc.2015.06.003.

Straus, S.E., & McAlister, F.A. (2000). Evidence-based medicine: a commentary on common criticisms. *Canadian Medical Association Journal*, *163*(7), 837–841.

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

Teira, D. (2011). Frequentist versus bayesian clinical trials. In Gifford, F. (Ed.) *Handbook of Philosophy of Medicine* (pp. 255–298). Wiley.

Teira, D., & Reiss, J. (2013). Causality, impartiality and evidence-based policy. In Mechanism and causality in biology and economics, (pp. 207–224). Springer.

Tillman, R.E., & Eberhardt, F. (2014). Learning causal structure from multiple datasets with similar variable sets. *Behaviormetrika*, *41*(1), 41–64. doi:10.2333/bhmk.41.41.

Unruh, W.G. (2008). Dumb holes: analogues for black holes. *Philosophical Transactions of The Royal Society A*, *366*, 2905–2913. doi:10.1098/rsta.2008.0062.

Upshur, R. (1995). Looking for rules in a world of exceptions: reflections on evidence-based practice. *Perspectives in Biology and Medicine*, *48*(4), 477–489. doi:10.1353/pbm.2005.0098.

Vandenbroucke, J.P., Broadbent, A., & Pearce, N. (2016). Causality and causal inference in epidemiology: the need for a pluralistic approach. *International Journal of Epidemiology*. doi:10.1093/ije/dyv341.

Waters, K.C. (2007). Causes that make a difference. *Journal of Philosophy*, *104*(11), 551–579. http://www.jstor.org/stable/20620058.

Weatherall, S., Ioannides, S., Braithwaite, I., & Beasley, R. (2014). The association between paracetamol use and asthma: causation or coincidence? *Clinical et Experimental Allergy*, *45*, 108–113. doi:10.1111/cea.12410.

Weber, M. (2006). The central dogma as a thesis of causal specificity. *History and Philosophy of the Life Sciences*, *28*(4), 595–609. http://www.jstor.org/stable/23334188.

Weed, D.L. (2005). Weight of evidence: a review of concept and methods. *Risk Analysis*, *25*(6), 1545–1557. doi:10.1111/j.1539-6924.2005.00699.x.

Weisberg, J. (2015). You've come a long way, bayesians. *Journal of Philosophical Logic*, *44*(6), 817–834. doi:10.1007/s10992-015-9363-9.

Wheeler, G., & Scheines, R. (2013). Coherence and confirmation through causation. *Mind*, *122*(485), 135–170. doi:10.1093/mind/fzt019.

Wimsatt, W.C. (1981). Robustness, reliability and overdetermination. In Brewer, M., & Colllins, B. (Eds.) Scientific inquiry and the social sciences: festschrift for Donald Campbell, (pp. 125–163). Jossey-Bass Publishers.

Wimsatt, W.C. (2012). Robustness, reliability, and overdetermination (1981). In Soler, L., Trizio, E., Nickles, T., & Wimsatt, W. (Eds.) Characterizing the robustness of science, boston studies in the philosophy of science, (Vol. 292 pp. 61–87): Springer, DOI doi:10.1007/978-94-007-2759-5_2.

Woodward, J. (2003). *Making things happen: a theory of causal explanation* (Oxford Studies in the Philosophy of Science). Oxford University Press.

Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology*, *13*(2), 219–240. doi:10.1080/13501780600733376.

Woodward, J. (2010). Causation in biology: stability, specificity and the choice of levels of explanation. *Biology and Philosophy*, *44*, 267–318. doi:10.1007/s10539-010-9200-z.

Worrall, J. (2007a). Evidence in medicine and evidence-based medicine. *Philosophy Compass*, *2*(6), 981–1022. doi:10.1111/j.1747-9991.2007.00106.x.

Worrall, J. (2007b). Why there's no cause to randomize. *British Journal for the Philosophy of Science*, *58*(3), 451–488. doi:10.1093/bjps/axm024.

Worrall, J. (2010). Do we need some large, simple randomized trials in medicine? In Suárez, M., Dorato, M., & Rédei, M. (Eds.) EPSA Philosophical issues in science: Launch of the European Philosophy of Science Association (pp. 289–301). doi:10.1007/978-90-481-3252-2_27.

Xie, L., Li, J., Xie, L., & Bourne, P. (2009). Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. *PLos Computational Biology*, *5*(5), 1–12. doi:10.1371/journal.pcbi.1000387.