Drug Saf (2015) 38:1–11
DOI 10.1007/s40264-014-0249-5

CURRENT OPINION

Causal Assessment of Pharmaceutical Treatments: Why Standards of Evidence Should not be the Same for Benefits and Harms?

Barbara Osimani · Fiorenzo Mignini

Published online: 18 December 2014

Abstract It is increasingly acknowledged both among epidemiologists and regulators that the assessment of pharmaceutical harm requires specific methodological approaches that cannot simply duplicate those developed for testing efficacy. However, this intuition lacks sound epistemic bases and delivers ad hoc advice. This paper explains why the same methods of scientific inference do not fare equally well for efficacy and safety assessment by tracing them back to their epistemic foundations. To illustrate this, Cartwright's distinction into clinching and vouching methods is adopted and a series of reasons is provided for preferring the latter to the former: (1) the need to take into account all available knowledge and integrate it with incoming data; (2) the awareness that a latent unknown risk may always change the safety profile of a given drug (precautionary principle); (3) cumulative learning over time; (4) requirement of probabilistic causal assessment to allow decision under uncertainty; (5) impartiality; and (6) limited and local information provided by randomised controlled trials. Subsequently, the clinchers/vouchers distinction is applied to a case study concerning the debated causal association between paracetamol and asthma. This study illustrates the tension between implicit epistemologies adopted in evaluating evidence and causality; furthermore, it also shows that discounting causal evidence may be a result of unacknowledged low priors or lack of valid alternative options. We conclude with a presentation of the changing landscape in pharmacology and the trend towards an increased use of Bayesian tools for assessment of harms.

B. Osimani (✉) · F. Mignini
University of Camerino, Camerino, MC, Italy
e-mail: barbara.osimani@unicam.it

Key Points

Causal assessment and scientific inference in pharmacology: inductivist approaches fare better than hypothetico-deductive ones when dealing with harms

Evidence standards for assessing benefits and harms should not be the same

A methodological framework for probabilistic (vs. categorical) causal assessment is needed

1 Introduction

Causal assessment for pharmaceutical harms is somewhat pressured by criteria that were mainly developed to evaluate efficacy and are therefore less suited to address the different problems associated with safety assessment. It is increasingly acknowledged both among epidemiologists and regulators that assessment of harms requires specific methodological approaches that cannot be parasitic on those developed for testing efficacy [1–9]. However, this intuition lacks sound epistemic bases and delivers ad hoc advice. We present here two contending paradigms of scientific inference and their epistemological rationales: one is the hypothetico-deductive method underlying the frequentist methodology of hypothesis testing and classical statistical inference—which informs the criteria underpinning evidence hierarchies as developed and advocated by evidence-based medicine—the other comprises the family of inductive approaches to scientific inference such as Bayesian hypothesis confirmation or inference to the best explanation/Peircean abduction. By drawing on Cartwright's distinction into clinching and vouching methods of scientific inference, we present such contending paradigms in connection with a case study that illustrates the tension between diverse epistemologies underpinning the alternative positions on the plausibility of a causal link between paracetamol and asthma. We show that vouching methods fare better when dealing with safety issues for a series of epistemic and pragmatic as well as methodological reasons. We also suggest that concern for confounding and other study quality issues may unknowingly hide a low prior for the hypothesis of causal link or practical concerns regarding the lack of valid alternative options. More generally, the two distinctive stances are at the origin of much talking across each other by the different scholars, in that methodological concerns are just the battlefield of opposite epistemological stances: a categorical one that needs conclusive evidence vs. a probabilistic one that also works with the available, albeit inconclusive, data.

2 Causal Inference and Evidence Standards

Causal inference brings together two strands of research programs: semantic and metaphysical inquiry on causality on one side (conceptual analysis and ontological status), and epistemologic and methodologic investigation of scientific inference on the other. The first strand of research dates back to Aristotle and his fourfold partition of the causal concept into efficient, teleologic, material and formal cause [10]. The contemporary debate on causality may be briefly subsumed under two main research projects: (1) identify the necessary and sufficient conditions for defining causality (metaphysical/semantic project); and (2) identify (perfect) indicators of causality to distinguish authentic from spurious causes (epistemological/methodological project). Obviously, the two strands are interconnected (and sometimes also confused).[1] For the present purpose, suffice it to say that necessary and sufficient conditions for causation have generally been cast in either counterfactual, probabilistic or manipulationist terms and that recommendations for distinguishing spurious from authentic causes have built on a mix of practical/methodological constraints and ontological claims that identify causes either with powers, counterfactuals or instantiation of law-like statements.

The second strand of research, foundations of scientific inference, has attracted the interest of philosophers since the emergence of modern science from what was previously known as "philosophy of nature". Galileo Galilei is universally acknowledged for having laid down the principles of modern experimental methodology [13]. Full-time philosophers such as Frances Bacon and John Stuart Mill have subsequently developed an analysis of scientific methodology by characterising it as an inductive procedure in contrast to deductive non-ampliative reasoning [14, 15]. At present, scientific reasoning is categorised under three main subheadings: inductive reasoning from empirical data to theory, hypothesis falsification through modus tollens [16],[2] and explanatory reasoning [19, 20]. Within statistics, two main approaches have emerged, one derives from Popperian falsificationism and proposes a series of procedures for rejecting or accepting hypotheses in a dichotomist fashion (Fisher hypothesis testing, and with some caveats, the Neyman Pearson approach); the other assigns hypotheses a degree of confirmation based on probability measures and takes into account all available evidence (various types of Bayesian statistics: more on this in the last section).

Considering the above, one could imagine that a significant number of different methods for causal assessment should be available by simply combining different approaches to scientific inference and different ways to conceptualise causality. Yet, at least in medical research, the preponderant standard is frequentist hypothesis testing. This is the result of a Popperian approach to scientific inference and an (oversimplified) counterfactual interpretation of causality. In classical hypothesis testing, the experimental outcome is expressed as the probability (p value) of observing the obtained result or more "extreme" results in the sample space if the treatment makes no difference (null hypothesis). By rejecting the null hypothesis, very low p values end up by corroborating the hypothesis of interest, i.e. the causal association between treatment and effect. Causal inference from the observed result relies thus on the assumption that the difference between the comparison groups is due to the contribution of the investigated factor (the treatment), and only to it. Blinding, intervention and randomisation are essential instruments in warranting this assumption [20–22] and evidence hierarchies are based on such warrants of internal validity.

[1] So, for instance, the potential outcome approach that infers causality from statistical data, is often presented in counterfactual terms; however, neglecting the gap between the metaphysics on which the counterfactual account of causation is based [11] and the complex ontological structure of the studied populations (see for instance [12]).

[2] Notwithstanding its seemingly rigorous logic, the hypothetico-deductive method is affected by the notorious Duhem–Quine problem: experimental evidence can never test a single hypothesis in isolation since this cannot be insulated from the theoretical system in which it is embedded and the inevitable experimental/methodological assumptions. Hence refuting evidence will reject the theoretical-experimental framework in toto without providing you with the means to discern what is true from what is false in it. Robustness analyses are intended as a means to overcome this problem [17, 18].

The role of randomisation within this picture is threefold: (1) to balance the proportion of possibly confounding factors (especially unknown ones) between treatment and control groups, so as to experimentally isolate the (distinctive contribution of) the cause under investigation (in analogy to the Galileian experimental method); (2) minimise (self-)selection bias (systematic error); (3) furthermore, repeated randomisation, would allow the true mean difference between treated and untreated sample populations to be approached in the limit (in the long run) [23, 24]. This third aspect concerns the estimation of the true effect size of the treatment by reducing random error. Although repeated sampling is never accomplished (for ethical and practical reasons),[3] meta-analyses (pooling data of diverse trials), or sampling as large populations as possible are considered the best feasible means to achieve the same goal. In this framework, the more likely a method is to exclude confounders and systematic/random errors, the more reliable is the inference we base on it; and because randomisation is supposed to bring a distinctive contribution in this respect; methods that use randomisation are ranked higher in evidence hierarchies all else being equal.[4]

Philosophers of science have animatedly discussed the privilege accorded to "randomised evidence". In particular, criticisms have been levelled against the evidence-based paradigm and its method of ranking evidence by essentially relying on randomisation [12, 21, 24, 27, 28, 29]. However, none of these contributions expressly addresses the specific challenges arising in causal assessment for harm. Epidemiologists such as Vandenbroucke as well as Ioannidis and colleagues [3, 4] have recognised the distinctive virtues and drawbacks of randomisation in assessing efficacy vs. harm. These suggestions have noteworthy implications when considering current emphasis on evidence hierarchies, because they imply an asymmetry in the way evidence of benefits and risks of health technologies should be evaluated.

Undeniably, current evidence standards are gradually starting to acknowledge the distinctive challenges posed by efficacy vs. safety assessment. Until very recently, CEBM levels of evidence, the evidence hierarchies developed by the Centre for Evidence Based Medicine at Oxford University [30], only distinguished between therapy, prognosis, diagnosis and economic analysis but failed to discriminate efficacy and harm assessment. Instead, the latest guidelines draw a distinction not only between evaluation of benefit and evaluation of harms, but also between common and rare harms. Still, privilege is accorded to evidence coming from systematic reviews of randomised controlled trials (RCTs), thereby obscuring the contribution of other types of evidence and especially the role of evidence amalgamation in safety issues. Similarly, Guyatt and colleagues [31] admit the difficulties inherent in the evaluation of evidence for harm, but propose a framework (the GRADE System) where evidence quality for assessment of harms follows the same criteria developed for efficacy evaluation. They admit that "patients vary widely in their preintervention or baseline risk of the adverse outcomes" ([32], 1,295) but no other advice follows from this observation than that of paying attention to the differing informative value of absolute vs. relative risk differences in subpopulations for the identification of differing patient characteristics. The reason why such suggestions seem to miss the point is that they stick to the requirements and constraints imposed on causal inference by statistical hypothesis testing while failing to see that there might be alternatives. More importantly, such suggestions fail to be grounded on a sound epistemic basis and seem rather ad hoc, although intuitively persuasive.

3 Towards a Pluralistic Methodology for Causal Assessment

Guidelines for drug approval, suspension and withdrawal heavily rely on frequentist statistics rationales, and this contributes to a sort of methodological monism, with the following consequences:

1. *Categorical causal assessment*. Such a paradigm is fundamentally categorical: hypotheses are either rejected or not, with no degrees in-between. The need for outright rejection/acceptance may be dropped in favour of methods that allow hypotheses to be assessed in a probabilistic fashion. Although this may seem detrimental at first glance, indeed it is extremely important in the case of cumulative knowledge accrual, as in the case of risk discovery, where "interim" causal assessments are required as soon as new information on previously undetected risks becomes available.
2. *One indicator of causality*. Causality is inferred from the difference between treated and untreated groups in the sample population; yet there may be available other types of indicators, which, jointly or separately, support causal inference; but standard methods are unable to exploit them fully because they do not relate to any formal method of causal inference.

In general, it can be said that this state of affairs imposes rigid constraints on the type of evidence that is allowed to

[3] Repeated randomisation is also physically impossible given that once a subject has received the treatment, in the next randomisation round she would not be the "same" subject as before [21].

[4] This also explains why case reports and observational data are considered sufficient evidence for causal claims to the extent that possible confounders/errors can be confidently excluded [4, 25, 26].

inform drug decisions and policies. This can be fully understood given that this methodology has been developed for testing efficacy; but it turns out that according to such standards, much evidence for harms must be considered of a lower-level quality and hence runs the risk of being systematically underestimated. Furthermore, notwithstanding the increasing awareness that "lower-level evidence" is a valid source of information for the risk profile of medications [26, 33, 34] current practices have difficulty in assigning it a precise epistemic status. A possible solution to this state of affairs is to enrich the methodological toolkit available by adopting a broader epistemological perspective. In her recent work, Nancy Cartwright has proposed an alternative to current evidence standards by distinguishing between two sorts of scientific methods: "clinchers" and "vouchers" [35]. These two typologies fundamentally rely on hypothetico-deductive (e.g. statistical hypothesis testing, see above) vs. inductive epistemologies. According to her account, the former are methods that provide you with a sure-fire guarantee (or something very close to it) about the validity of your inference; the latter instead can only "vouch for" it, and can be refuted by defeating evidence. "Vouchers" do not definitely refute or prove a given hypothesis, instead they aim to provide a way to assess the degree of plausibility of a given hypothesis given a set of data and, possibly, to measure it probabilistically. The two main approaches within this paradigm are Bayesian methods of hypothesis confirmation and inference to the best explanation (also related to Peircean "abduction").

It must be noted that a similar view was also expressed by Sir Austin Bradford Hill 50 years ago in his President's Address [36] inaugurating the Section of Occupational Medicine of the Royal Society of Medicine; that is, a discipline mostly concerned with exposure to hazards. After presenting his nine guidelines for detecting and assessing causal relationships he claims: "None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a sine qua non. What they can do, with greater or less strength, is to help us make up our minds in the fundamental question—*is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect?*" (emphasis added). Thus, Hill both refers to explanatory power as well as to hypothesis likelihood as reliable grounds to justify causal judgements on different grounds, and explicitly presents his approach as an alternative to frequentist hypothesis testing. More than that, Hill points to a reversal of perspective when dealing with hazards: in evaluating harms you generally do not start with a specific hypothesis and then collect evidence to test it, but rather the reverse: you happen to observe evidence, which you are called on to explain in some way.

Hill's justification of causal claims depends on heterogeneity of methods and evidence. The reasoning behind it may be related to Campbell's notion of "triangulation" or "robustness": [17, 18] because it is unlikely that independent pieces of evidence all confirm a given hypothesis H, if it is not true, so independent types of evidence pointing to the same hypothesis have a high confirmatory power. Evidence at different levels (e.g. molecular, clinical, epidemiological) jointly works in an analogous way. Hence, Hill proposes a list of indicators of causality (biological plausibility, strength and specificity of the association) that may cumulatively contribute to support the causal association. This sharply differentiates his approach from that developed by evidence guidelines (for efficacy assessment), which present a list of indicators of evidence quality. The two approaches are orthogonal because the latter is focused on quality of evidence, whereas the former is focused on evidence amalgamation, but they use different tools for the same purpose: justification of causal claims. The point is that they do not fare equally well with regard to assessment of harms. Whereas inductive approaches can use also evidence coming from hypothetico-deductive methods of scientific inference, such as RCTs and incorporate it in the overall hypothesis assessment, the reverse does not hold.

4 Causal Assessment of Harm vs. Benefit: Epistemic and Pragmatic Asymmetries

By regimenting efficacy and safety assessment with the same standards, we fail to pay heed to important epistemic and pragmatic asymmetries. Rudén and Hansson [37] observe that the focus of research in risk detection is on false negatives, rather than on false positives; in the sense that, owing to epistemic and pragmatic reasons—the approval procedure is focused on testing efficacy, whereas risk detection is rather delegated to postmarketing monitoring—as well as to the influence of conflict of interests, the former type of error is more probable than the second. Hence, the main issue is failure to see causation, rather than discerning spurious from authentic causes. This epistemic asymmetry explains the claim by Papanikolaou et al. [3] to the effect that: 'it may be unfair to invoke bias and confounding to discredit observational studies as a source of evidence on harms' (p. 640, emphasis added). When dealing with efficacy, observational studies are not considered enough to warrant causal claims by default, for fear of fraud related to false positives and obvious conflict of interests in exploiting them; however, the opposite is valid for harms where positives (whether false or true) are not desired and therefore tend to be discounted (the history of pharmaceutics offers a series of examples in this sense:

from the thalidomide tragedy, to the Cronassial© case, Vioxx© and Bextra©). Hence, using the argumentation of bias and confounding in the two settings grounds on opposite concerns.

To this type of asymmetry one should also add that, when dealing with harms, the problem of external validity is "reversed", in three senses: (1) in the case of unintended/unexpected effects, the information searched for is not whether the target population will experience the same outcomes observed in the study population, but whether it will experience additional outcomes that have not been detected during the study; (2) if there is evidence from a study, of whatever type, that a drug is causally linked to an adverse event, then one already knows that this effect can occur, and because the drug harms "somewhere" it may harm somewhere else too: this constitutes causally relevant information without any need of further external validity warrant; (3) information about risk subgroups may be used to avoid/minimise harm, for instance by adopting preventive measures (e.g. the use of Mesna to reduce the incidence of haemorrhagic cystitis and haematuria when a patient receives ifosfamide or cyclophosphamide for cancer chemotherapy); by excluding some patient groups (e.g. hypersusceptibility to abcavir in patients with the HLA B*5701 allele); or by monitoring for adverse reactions (e.g. risk of neutropenia in patients treated with clozapine).[5] This is a mirror image to how information about causal interaction is used in relation to benefits, i.e. to detect specific patient subgroups where efficacy is enhanced, and thus to promote treatment use among them. The different concerns affecting assessment of harms, with respect to the assessment of efficacy, provides a series of reasons for preferring vouchers to clinchers when dealing with harm.

4.1 Integration of Prior Knowledge and Observation

Frequentist statistics does not allow one to incorporate priors in hypothesis evaluation. This is particularly detrimental in the case of harm assessment considering that much knowledge of the drug behaviour may be inferred analogically from same-class molecules. Furthermore, most compounds are characterised by promiscuity, meaning that they bear some affinity to off-target proteins: this is at the origin of most adverse reactions at the clinical level and integrating information about biochemical constraints, molecular mechanisms and biological pathways at the systems level considerably enhances the predictability of drug–organism interactions [38]. More generally, theoretical knowledge at different levels may be fruitfully combined (and amalgamated with available empirical evidence) to anticipate risk [39]

4.2 High Default Prior for an Undefined Risk

Although when a drug is approved for marketing, many of its unintended (possibly harmful) effects are unknown, there is widespread awareness that a latent unspecific risk is associated with it. This is the rationale behind the introduction of the precautionary principle in the pharmaceutical domain and is also reflected in the regulation that introduced the notion of "development risk" (or "potential risk"), as well as the pharmacosurveillance system [40].

4.3 Cumulative Learning and the Virtues of Probabilistic vs. Categorical Causal Assessment

Evidence accumulates over time and there comes a point where the signal strongly suggests causation without demonstrating it. An epistemology grounded on hypothesis rejection is of no help in this situation. Indeed, following the precautionary principle, you are not supposed to wait for "scientific proof" of causal connection, but rather to act as soon as its probability is high enough to recommend countermeasures in relation to the risk–benefit balance (see next point) [41]. Hence, what is needed is an instrument that allows one to make decisions under uncertainty.

Pharmaceutical risk management and decision making follows analogous criteria to those developed for standard health technology assessment, i.e. evaluation of costs and benefits latu sensu. This means that pharmaceutical products are kept in the market insofar as the expected benefit outweighs the expected harm. However, a problem arises when a harmful effect is only suspected to be associated with the drug but a causal connection between them has yet not been established. Following the precautionary principle, hypotheses of causal relationships need not be rejected or accepted: it is sufficient that they are strong enough with respect to the risk that is associated with the technology under examination [42]. This change of paradigm in administrative and tort rule, where in principle causation needs to be firmly established for assigning culpability, has been mainly fostered by environmental law, but has been anticipated by the German legislation on pharmaceuticals as a result of the pressure caused by the Contergan case (thalidomide) and the related sentence [41].

4.4 Impartiality

Impartiality assumes in the case of efficacy vs. safety assessment opposite characteristics. Efficacy must be tested against fraud (which explains the success of randomised trials in pharmaceutical regulation) [24]. As for harms,

[5] We thank an anonymous reviewer for suggesting us point 2 and 3 of the list, examples included.

fraud is linked to holding back safety information (see the Vioxx case), [40] hence the point of contention is reversed. Teira [24] conceptualises impartiality as a way to deal with uncertainty such that it cannot be exploited by some party's private interest. Waiting for an RCT to definitively prove that an observed risk is really associated with a suspected drug, exactly represents the case in which the uncertainty about the causal association may be exploited by the industry's financial interests, to the detriment of patient safety and public health expenditures.

4.5 Limited Causal Information of RCTs

However useful, RCTs deliver limited and purposely decontextualised information; the statistical principles on which they are grounded have been developed to test fertilisers on plant growth.[6] The causal structure here is much closer to physical causality [44]—plots of lands do not react to fertilisers in the same way as humans absorb and metabolise drugs (pharmacokinetics, pharmacodynamics)—and is not as rich in feedback loops, threshold effects, interactive causality and system level effects as it is characteristic of complex biological organisms, where such phenomena are the rule [45]. Yet, because different types of populations may experience different effects by taking the same drug, conducting larger and larger RCTs or pooling data in meta-analyses would not reach the purpose. Indeed, for results to be at all meaningful, the studies included in the meta-analysis need to be as homogeneous as possible: they need to have been conducted with the same inclusion-exclusion criteria, the same type of control and the same context of treatment administration. Methods to quantify heterogeneity, such as fixed-effect or random-effect analysis, aim to measure the study quality rather than to account for strata differences. The epistemic asymmetry between efficacy and safety assessment examined with reference to RCTs replicates also at the meta-level and is reflected in the different use of meta-analyses in the two settings. In efficacy assessment, they are used as a type of "robustness analysis": meta-analysis should confirm the (possibly conflicting) results of individual RCTs; instead concerning adverse events, they are explicitly used to detect risks for which individual RCTs are underpowered. Recent contributions to the methodology of systematic reviews also go in the direction to emphasising internal validity [46] by appropriately selecting studies, while obscuring the importance of other issues in evidence amalgamation, such as the combination of heterogeneous evidence for the purpose of "connecting the dots" between different constituents of a phenomenon.

The tension between clinching and vouching epistemologies is often left implicit behind methodological debates, where the tacit idea is not only that randomisation guarantees causation; but also that it is the only means to achieve such warrant. We present a case study concerning causal assessment of harm, where this tension becomes quite visible and has practical implications for healthcare and drug regulation.

5 Case Study: The Paracetamol Enigma in Asthma

There is an ongoing debate on the possible causal association between paracetamol and asthma, and its implications for prescription practice (especially in paediatrics). On one side, there are those who feel reluctant to accept this claim, on grounds that it is not rooted on randomised clinical trials [47–52]. Particularly, these authors express the concern that the paracetamol–asthma relationship may be explained by reverse causation, or confounding by indication. Other authors are less imperative on the matter but equally require or recommend the performance of adequately powered placebo-controlled trials to establish causation [53, 54]. On the other side, we observe an alternative approach to causal assessment and evidence evaluation: Beasley et al. [55] assert for instance that "when the study findings are considered together with other available data, there is substantive evidence that paracetamol use in childhood may be an important risk factor for the development and/or maintenance of asthma" (p. 1,570, emphasis added). An even stronger commitment to the hypothesis of causal association is expressed by McBride [56] who, considering all the evidence available (see Table 1), claims that evidence of causal association can by now be regarded as strong enough to warrant a change in prescription practice.

More recently, Martinez-Gimeno and García-Marcos, [73] emphasise that apart from tobacco smoke exposure, the association between paracetamol and wheezing disorders is the most robust among all other candidate factors, genetic or environmental. Furthermore, they are against the performance of double-blind RCTs with placebo, as "a placebo arm would be impractical and unethical, because it would subject participants to a substandard and unacceptable treatment during a very long time" (p. 114).

The dissent concerning the best course of action among scholars is ultimately caused by differing epistemological views, which are nonetheless left implicit. Those recommending the performance of placebo-controlled RCTs are in line with the rationales underlying evidence hierarchies, and hence the clinching epistemology associated with it.

[6] Randomised controlled trials have a long history that starts before their statistical formalisation by R. Fisher: their development in the history of experimentation is related to obtaining information that is directly action-guiding (vs. purely epistemic), see [43]. We thank an anonymous reviewer for clarifying this point.

Table 1 The consistency of interdisciplinary evidence makes the causal association between asthma and acetaminophen plausible enough to change prescription practice (in paediatrics): McBride (2011)

1. Strength of the association [57–59]
2. Robustness of association across geography, culture and age [51, 53–55, 60–63]
3. Dose–response relationship between acetaminophen exposure and asthma [58, 59]
4. Coincidence of time trends in acetaminophen use and asthma increase [64]
5. Lack of other equally strong causal explanations [65–67]
6. Relationship between asthma epidemic and per-capita sales of acetaminophen across countries [68]
7. Plausible mechanism [47, 63, 69–72]

Thus they insist on the elimination of any suspicion of confounding, especially confounding by indication [48, 53] before any causal claim can be established on firm grounds. On the other side, advocates of the causal link, especially McBride [56] and Beasley et al. [55], point to the joint support of different and independent sources of evidence as a valid basis for dropping any need for RCTs. This corresponds to the idea that amalgamation of heterogeneous evidence may be a valid alternative to hypothesis testing when safety signals[7] coming from different sources are suggestive of causal association even though they cannot conclusively prove it.

The case is however complicated by the fact that there are no safer alternative treatments available: various other non-opioid analgesics, such as metamizole and aminophenazone, are associated with severe adverse reactions. As for ibuprofen, we cannot be confident that its lower risk profile is just owing to its more recent history in the pharmaceutical market. The question though is that such practical constraints influence the causal assessment concerning whether indeed paracetamol causes an increased risk in asthma incidence or prevalence. Therefore, the risk of confounding may be emphasised because of a concern about the lack of reasonable alternatives, rather than because of a real epistemic issue.

According to Martinez-Gimeno and García-Marcos [73] for instance, prior knowledge about the molecule, considered a harmless analgesic, explains in this case the reluctance to accept the causal hypothesis between paracetamol and asthma. The implicit reasoning behind the reticence to see some harm as drug related is thus due to prejudicial priors of innocuousness for a given molecule, which may deceptively induce medical scientists as well as health practitioners and the responsible authorities to discount the drug as a possible cause for observed safety signals. Although apparently speculative, this suggestion is motivated, at least in this case, by considering the preponderance of evidence in favour of causation rather than against it. In fact, according to the authors, the quantity, variety and cross-confirmatory value of epidemiological evidence speaks so strongly in favour of the causal hypothesis (see Table 1), that the reluctance of other scholars to agree on the causal association can be explained in their opinion only by presuming that it is downplayed because of a "prejudice" fostered by premarketing data showing paracetamol to be relatively harmless.

Hence, the declared concern for confounders would rather hide a conservative low prior for harmfulness. This means that instead of explicitly taking prior knowledge into account, this would be allowed to influence the interpretation of observational evidence "in an underhand manner".

Indeed, an alternative explanation is also possible, i.e. that scholars who do not accept such causal association tend to think categorically rather than probabilistically and so, even in the case where they would be inclined to give credit to such a hypothesis, nevertheless they would require conclusive evidence before they can definitely accept it, and rightly so (according to their perspective). However, this would run against the precautionary approach recommended in risk prevention and minimisation, which prompts the undertaking of adequate interventions even in the absence of incontrovertible proofs of causal association.

The lesson here, however, is that relating such methodological disputes to their epistemological underpinnings casts some light on why scholars seem to be talking across each other at some point of the dispute: adherents to vouching methods are satisfied by less than conclusive evidence and use it for decisions, whereas adherents to clinching methods require conclusive evidence before pronouncing themselves in favour of a causal association. Both might think that the issue is just methodological (relating to confounders or bias), while instead the problem lies at a deeper level; i.e. whether one adheres to probabilistic or categorical thinking.

The paracetamol case illustrates that evidence for causal assessment of harms may be disqualified on several grounds, which may not be cogent and may be motivated by epistemological stances that are left implicit. The temporal dynamics of the case also show that cumulative learning around the causal association between paracetamol and asthma emerges and grows through evidence of different types, which demand to be accounted for to provide guidance for action even before a definitive proof is available. The standard clinching way to assess causality

[7] Contrary to commonsense intuition, "safety signal" does not refer to indicators of safety, but rather the contrary, i.e. signals suggesting possible risks associated with the drug.

does not have the tools to carry out this task, in that this requires a combination of methods and a qualitative/quantitative integration of heterogeneous data. Adopting a "vouching approach" for assessing harms is therefore the answer to the problem, although still much work is needed to develop a system of evidence evaluation grounded on this basis.

6 Evidence Standards in Evolution

Guidelines on pharmacosurveillance and signal detection promote a flexible strategy to safety studies and encourage an "information-based" rather than a "power-based" approach to causal assessment (see the guidelines for good pharmacovigilance practice of the European Medicines Agency - Heads of Medicines Agency). Furthermore, most drug withdrawals are based on individual cases or case series reporting dramatic/fatal effects [74–76]; but for less dramatic outcomes, such as the asthma case cited above, no clear guidance is available and adopting standards of evidence developed for testing efficacy may not be an adequate strategy. It should also be considered that anecdotal reports constitute about 30 % of the world literature on adverse drug reactions, while systematic reviews constitute less than 3 % [77].[8] Much of the evidence for harms comes from anecdotal reports, case series, or survey data. The role of this "lower-level" evidence is increasingly acknowledged to be a valid source of information that contributes to assessing the risk profile of medications on theoretical [26, 33, 34] or empirical grounds [78–80], but current practices have difficulty in assigning a precise epistemic status to this type of evidence and integrating it with more standard methods of hypothesis testing.

A recent methodological review [9] authored by the Drug Information Association Bayesian Scientific Working Group, points to the unique challenges faced by safety evaluation in drug development and surveillance: "unlike efficacy assessments that are primarily driven by hypotheses, safety assessments involve a wide range of safety measures (such as adverse events, laboratory, etc.), which need to be studied together to make an overall safety conclusion" ([9], 15). This quote mirrors the considerations expressed by Hill concerning the distinctive features of evidence for hazards with respect to efficacy. Indeed, the Bayesian approach can be considered in this sense a vouching method in contrast to the standard frequentist approach underpinning RCTs and evidence-based medicine guidelines. The advantage of Bayesian statistics over frequentist methods is that it allows the assigning of a probability measure to the hypothesis of causal link, thereby allowing decisions under uncertainty. Furthermore, Bayesian approaches to scientific inference may incorporate diverse kinds of evidence, beyond strictly statistical data. The review [9] supports Bayesian methods for the design and analysis of safety trials because of their ability to incorporate historical (heterogeneous) knowledge in the prior, to adapt sample size on the basis of accruing knowledge, and generally to uncover problems at an earlier stage, which practically means less false negatives and hence increased risk prevention.

A Bayesian approach to safety assessment is not only beneficial with regard to these dimensions (and related ethical issues), but, if consistently adopted for the entire product life cycle, would undeniably represent an important turn for safety assessment in terms of efficiency and epistemic warrant. In fact, by allowing multiple sources of evidence to be incorporated in the probability function associated with the hypothesis under investigation, the Bayesian approach optimises the use of the available evidence; furthermore, by using explicit priors and separating them from the likelihood function, it lets researchers and other stakeholders see the impact of prior knowledge (theoretical or from other sources, such as past records) on the final study result in an unequivocal and transparent way. Yet, there is a double standard concerning safety assessment in pharmacology: whereas Bayesian tools for "pattern recognition" encounter no objections for the purpose of signal detection (see for instance the work of the Uppsala Monitoring Center), still the implicit standard in safety assessment and decisions concerning drug withdrawal relies on the same criteria developed for efficacy assessment, characterised by a stringent canon of statistical methods based on standard hypothesis testing. The evidence-based medicine emphasis on internal validity criteria, borrowed from classical statistical methodology, is so entrenched that it may obscure the different challenges posed by safety evaluation with respect to testing efficacy. In the end, a too rigid attitude towards evidence quality may run against the reasons for which quality standards have been introduced.

7 Conclusion

Causal assessment for harms is somewhat pressured by criteria that were mainly developed to evaluate efficacy and are therefore less suited to address the different problem of detecting and assessing harm. We presented here two contending paradigms of evidence evaluation and made their epistemological rationales explicit: one is the hypothetico-deductive method underlying the frequentist methodology of hypothesis testing and classical statistical inference, which also informs the criteria underpinning

[8] We thank an anonymous reviewer for signaling us these figures.

evidence hierarchies as developed and advocated by evidence-based medicine. The other comprises the family of inductive approaches to scientific inference such as Bayesian hypothesis confirmation and inference to the best explanation/Peircean abduction. These approaches have been mapped into Carwright's distinction into clinching and vouching methods and a series of reasons have been provided as grounds for preferring the latter to the former when dealing with harms. These are related to the fact that, when dealing with harms, false negatives, rather than false positives constitute the main concern. We also presented a case study, the debated causal association between paracetamol and asthma, which illustrates the tension between such different epistemological stances and concluded with a presentation of the changing landscape in drug safety assessment, where Bayesian methods are gaining increasing ground. This is because of their ability to optimise the use of available evidence by incorporating historical (heterogeneous) 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.

Funding and conflicts of interest No sources of funding were used to assist in the preparation of this study. Barbara Osimani and Fiorenzo Mignini have no conflicts of interest that are directly relevant to the content of this study.

References

1. Vandenbroucke JP, Psaty BM. Benefits and risks of drug treatments: how to combine the best evidence on benefits with the best data about adverse effects. JAMA. 2008;300(20):2417–9.
2. Psaty B, Vandenbroucke JP. Opportunities for enhancing the FDA guidance on pharmacovigilance. JAMA. 2008;300(8):952–3.
3. Papanikolaou PN, Christidi GD, Ioannidis JPA. Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. CMAJ. 2006;174(5):635–41.
4. Vandenbroucke JP. Observational research, randomised trials, and two views of medical science. Plos Med. 2008;5(3):339–43 (quotation in the text are from the longer version: 1–28).
5. European Medicines Agency. Heads of Medicines Agencies. Guideline on good pharmacovigilance practices (GVP). Module VII. Periodic safety update report. 2012. EMA/81692/2011.
6. European Medicines Agency. Heads of Medicines Agencies. Guideline on good pharmacovigilance practices (GVP). Module VIII. Post-authorization safety studies. 2012. EMA/813938/2011.
7. European Medicines Agency. Heads of Medicines Agencies. Guideline on good pharmacovigilance practices (GVP). Module IX. Signal management. 2012. EMA/827661/2011.
8. European Medicines Agency. Heads of Medicines Agencies. Guideline on good pharmacovigilance practices (GVP). Module X. Additional monitorin. 2012. EMA/169546/2012.
9. Price KL, Xia HA, Lakshminarayanan M, Madigan D, Manner D, Scott J, Stamey JD, Thompson L. Bayesian methods for design and analysis of safety trials. Pharm Stat. 2014;13:13–24.
10. Aristotle, Metaphysics, Oxford: Oxford University Press, 1924, Books I, V.
11. Lewis D. Counterfactuals. Oxford: Blackwell Publishers, Cambridge: Harvard University Press, 1973, Reprinted with revisions; 1986.
12. Cartwright N. Are RCTs the gold standard? Biosocieties. 2007;2:11–20.
13. Drake S. Essays on Galileo and the history and philosophy of science. In: Swerdlow NM, Levere TH, editors, vol. 3. Toronto: University of Toronto Press; 1999.
14. Bacon F. Novum organum. In: Spedding J, Ellis RL, Heath DD, editors. The works. vol. 8, Boston: Taggard and Thompson; 1863.
15. Mill JS. A system of logic. New York: Harper & Brothers; 1882.
16. Popper K. The logic of scientific discovery. London: Routledge; 1992.
17. Wimsatt WC. Robustness, reliability and overdetermination. In: Brewer MB, Colllins BE, editors. Scientific inquiry and the social sciences: Festschrift for Donald Campbell. San Francisco, CA: Jossey-Bass; 1981. p. 125–63.
18. Wimsatt WC. Robustness: material, and inferential, in the natural and human sciences. In: Soler L, Trizio E, Nickles T, Wimsatt WC, editors. Characterizing the robustness of science. Berlin: Springer; 2012. p. 89–103.
19. Peirce CS. Lecture two: types of reasoning. In: Ketner KL, editor. Reasoning and the logic of things, the Cambridge conference lectures of 1898. Cambridge: Harvard University Press; 1992. p. 123–42.
20. Lipton P. Inference to the best explanation. New York: Routledge; 2004.
21. Papineau D. The virtues of randomization. B J Phil Sci. 1993;45(2):437–50.
22. Worrall J. Why there's no cause to randomize. B J Phil Sci. 2007;58:451–88.
23. Basu D. Randomization analysis of experimental data: the fisher randomization test. J Am Stat Assoc. 1980;75(371):593–5.
24. Teira D. Frequentist versus Bayesian clinical trials. In: F. Gifford, editor. Handbook of the philosophy of science. Philosophy of Medicine. vol. 16, 2011. p. 255–298.
25. Glasziou P, Chalmers I, Rawlins M, McCullock P. When are randomized trials necessary? Picking signal from noise. BMJ. 2007;334:349–51.
26. Howick J, Glasziou P, Aronson JK. The evolution of evidence hierarchies: what can Hill's 'guidelines for causation' contribute? J R Soc Med. 2009;102:186–94.
27. Worrall J. Do we need some large, simple randomized trials in medicine? EPSA philosophical issues in the science. Dordrecht: Springer; 2010. p. 289–301.
28. Papineau D. Metaphysics over methodology: or, why infidelity provides no grounds to divorce causes from probabilities. In: Galavotti MC, Suppes P, Costantini P, editors. Stanford: Stochastic Causality CSLI Publications; 2001. p. 15–38.
29. Pearl J. Causality. Models, reasoning, and inference. New York, Cambridge: Cambridge University Press; 2000.
30. Howick J, Chalmers I, Glasziou P,Greenhalgh T, Heneghan C, Liberati A, Moschetti I, Phillips B, Thornton H. The 2011 Oxford CEBM evidence levels of evidence (introductory document). Oxford Centre for Evidence-Based Medicine; 2011. http://www.cebm.net/index.aspx?o=5653. Accessed 8 Dec 2014.
31. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, DeBeer H, Jaeschke R, Rind D, Meerpohl J, Dahm P, Schünemann HJ. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383–94.
32. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, Glasziou P, Jaeschke R, Akl EA, Norris S, Vist G, Dahm P, Shukla VK, Higgins J, Falck-Ytter Y,

Schünemann HJ, GRADE Working Group. GRADE guidelines: 7. Rating the quality of evidence: inconsistency. J Clin Epidemiol. 2011;64(12):1294–302.
33. Aronson JK, Hauben M. Anecdotes that provide definitive evidence. BMJ. 2006;333(16):1267–9.
34. Hauben M, Aronson JK. Gold standards in pharmacovigilance: the use of definitive anecdotal reports of adverse drug reactions as pure gold and high-grade ore. Drug Saf. 2007;30(8):645–55.
35. Cartwright N. The art of medicine: A philosopher's view of the long road from RCTs to effectiveness. Lancet. 2011;377:1400–1.
36. Hill AB. The environment and disease: association or causation? Proc R Soc Med. 1965;58:295–300.
37. Rudén C, Hansson SO. Evidence-based toxicology: "sound science" in new disguise. Int J Occup Environ Health. 2008;13(4):299–306.
38. Xie L, Li J, Xie L, Bourne PE. Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. PLOS Comput Biol. 2009;5(5):e1000387.
39. Bai JPF, Abernethy DR. Systems pharmacology to predict drug toxicity: integration across levels of biological organization. Annu Rev Pharmacol Toxicol. 2013;53:451–73.
40. Osimani B. Pharmaceutical risk communication: sources of uncertainty and legal tools of uncertainty management. Health Risk Soc. 2010;12(5):453–69.
41. Osimani B. The precautionary principle in the pharmaceutical domain: a philosophical enquiry into probabilistic reasoning and risk aversion. Health Risk Soc (Special Issue: Health Care through the Lens of Risk). 2013;15(2):123–43.
42. Scheu, G. In Dubio Pro Securitate. Contergan, Hepatitis-/AIDS-Blutprodukte, Spongiformer Humaner Wahn und kein Ende? Grundrechtliche Gefahrenvorsorge für Leib und Leben im Recht der Produkt- und Arzneimittelsicherheit: auch unter Aspekten der Europäisierung und Globalisierung. Nomos: Baden-Baden; 2003.
43. Hansson SO. Why and for what are clinical trials the gold standard? Scand J Public Health. 2014;41(Suppl 13):41–8.
44. Thompson RP. Causality, theories and medicine. In: Illari PM, Russo F, Williamson J, editors. Causality in the sciences. Oxford: OUP; 2011.
45. Smith SW, Hauben M, Aronson JK. Paradoxical and bidirectional drug effects. Drug Saf. 2012;35(3):173–89.
46. Waddington H, et al. How to do a good systematic review of effects in international development: a tool kit. J Dev Eff. 2012;4(3):359–87.
47. Eneli I, Sadri K, Camargo C, Barr RG. Acetaminophen and the risk of asthma: the epidemiologic and pathophysiologic evidence. Chest. 2005;127(2):604–12.
48. Allmers H, Skudlik C, John SM. Acetaminophen use: a risk for asthma? Curr Allergy Asthma Rep. 2009;9(2):164–7.
49. Johnson CC, Ownby DR. Have the efforts to prevent aspirin-related Reye's syndrome fuelled an increase in asthma? Clin Exp Allergy. 2011;42(3):296–8.
50. Karimi M, Mirzaei M, Ahmadieh MH. Acetaminophen use and the symptoms of asthma, allergic rhinitis and eczema in children. Iran J Allergy Asthma Immunol. 2006;5(2):63–7.
51. Wickens K, Beasley R, Town I, Epton M, Pattemore P, Ingham T, Crane J, New Zealand Asthma and Allergy Cohort Study Group. The effects of early and late paracetamol exposure on asthma and atopy: a birth cohort. Clin Exp Allergy. 2011;41(3):399–406. doi:10.1111/j.1365-2222.2010.03610.x (Epub 2010 Sep 29).
52. Chang KC, Leung CC, Tam CM, Kong FY. Acetaminophen and asthma: spurious association? Am J Respir Crit Care Med. 2011;183(11):1570.
53. Holgate ST. The acetaminophen enigma in asthma. Am J Respir Crit Care Med. 2011;183:147–8.
54. Henderson AJ, Shaheen SO. Acetaminophen and asthma. Paediatr Respir Rev. 2013;14(1):9–15.
55. Beasley RW, Clayton TO, Crane J, Lai CK, Montefort SR, Mutius EV, Stewart AW, ISAAC Phase Three Study Group. Acetaminophen use and risk of asthma, rhinoconjunctivitis, and eczema in adolescents: International Study of Asthma and Allergies in Childhood Phase Three. Am J Respir Crit Care Med. 2011;183(2):171–8.
56. McBride JT. The association of acetaminophen and asthma prevalence and severity. Prediatrics 2011;128. http://pediatrics.aappublications.org/content/128/6/1181.full.pdf+html?sid=29b845ce-0023-48dd-9835-3c87a6b45b69. Accessed 8 Dec 2014.
57. Shaheen SO, Potts J, Gnatiuc L, Makowska J, Kowalski ML, Joos G, van Zele T, van Durme Y, De Rudder I, Wöhrl S, Godnic-Cvar J, Skadhauge L, Thomsen G, Zuberbier T, Bergmann KC, Heinzerling L, Gjomarkaj M, Bruno A, Pace E, Bonini S, Fokkens W, Weersink EJ, Loureiro C, Todo-Bom A, Villanueva CM, Sanjuas C, Zock JP, Janson C, Burney P, Selenium and Asthma Research Integration project; GA2LEN. Selenium and Asthma Research Integration Project; GA2LEN. The relation between paracetamol use and asthma: a GA2LEN European case-control study. Eur Respir J. 2008;32(5):1231–6.
58. Barr RG, Wentowski CC, Curhan GC, Somers SC, Stampfer MJ, Schwartz J, Speizer FE, Camargo CA Jr. Prospective study of acetaminophen use and newly diagnosed asthma among women. Am J Respir Crit Care Med. 2004;169(7):836–41.
59. Lesko SM, Louik C, Vezina RM, Mitchell AA. Asthma morbidity after the short-term use of ibuprofen in children. Pediatrics 2002;109 (2). http://pediatrics.aappublications.org/content/109/2/e20.full.pdf+html. Accessed 8 Dec 2014.
60. Beasley R, Clayton T, Crane J, von Mutius E, Lai CK, Montefort S, Stewart A, ISAAC Phase Three Study Group. Association between paracetamol use in infancy and childhood, and risk of asthma, rhinoconjunctivitis, and eczema in children aged 6–7 years: analysis from phase three of the ISAAC programme. Lancet. 2008;372(9643):1039–48.
61. Etminan M, Sadtsafavi M, Jafari S, Doyle-Waters M, Aminzadeh K, Fitzgerald JM. Acetaminophen use and risk of asthma in children and adults: a systematic review and meta-analysis. Chest. 2009;136(5):1316–23.
62. Amberbir A, Medhin G, Alem A, Britton J, Davey G, Venn A. The role of acetaminophen and geohelminth infection on the incidence of wheeze and eczema: a longitudinal birth-cohort study. Am J Respir Crit Care Med. 2011;183(2):165–70.
63. Farquhar H, Stewart A, Mitchell E, Crane J, Eyers S, Weatherall M, Beasley R. The role of paracetamol in the pathogenesis of asthma. Clin Exp Allergy. 2010;40(1):32–41.
64. Varner AE, Busse WW, Lemanske RF Jr. Hypothesis: decreased use of pediatric aspirin has contributed to the increasing prevalence of childhood asthma. Ann Allergy Asthma Immunol. 1998;81(4):347–51.
65. Seaton A, Godden DJ, Brown K. Increase in asthma: a more toxic environment or a more susceptible population? Thorax. 1994;49:171–4.
66. Eder W, Ege JM, von Mutius E. The asthma epidemic. NEJM. 2006;355(21):2226–35.
67. Platts-Mills T, Vaughan J, Squillace S, et al. Sensitisation, asthma, and a modified Th2 response in children exposed to cat allergen: a population-based cross-sectional study. Lancet. 2001;357:752–6.
68. Newson RB, Shaheen SO, Chinn S, Burney PG. Paracetamol sales and atopic diseases in children and adults: an ecological analysis. Eur Respir J. 2000;16(5):817–23.
69. Nassini R, Materazzi S, Andrè E, Sartiani L, Aldini G, Trevisani M, Carnini C, Massi D, Pedretti P, Carini M, Cerbai E, Preti D, Villetti G, Civelli M, Trevisan G, Azzari C, Stokesberry S, Sadofsky L, McGarvey L, Patacchini R, Geppetti P.

Acetaminophen, via its reactive metabolite *N*-acetyl-*p*-benzoquinoneimine and transient receptor potential ankyrin-1 stimulation, causes neurogenic inflammation in the airways and other tissues in rodents. FASEB J. 2010;24(12):4904–16. doi:10.1096/fj.10-162438.

70. Graham NM, Burrell CJ, Douglas RM, Debelle P, Davies L. Adverse effects of aspirin, acetaminophen, and ibuprofen on immune function, viral shedding, and clinical status in rhinovirus-infected volunteers. J Infect Dis. 1990;162(6):1277–82.
71. Galindo PA, Borja J, Mur P, et al. Anaphylaxis to paracetamol. Allergol Immunopathol. 1998;26:199–200.
72. De Paramo BJ, Gancedo SQ, Cuevas M, et al. Paracetamol (acetaminophen) hypersensitivity. Ann Allergy Asthma Immunol. 2000;85:508–11.
73. Martinez-Gimeno A, García-Marcos L. The association between acetaminophen and asthma: should its pediatric use be banned? Expert Rev Respir Med. 2013;7(2):113–22.
74. Olivier P, Montastruc JL. The nature of the scientific evidence leading to drug withdrawals for pharmacovigilance reasons in France. Pharmacoepidemiol Drug Saf. 2006;15:808–12.
75. Arnaiz JA, Carne X, Riba N, Codina C, Ribas J, Trilla A. The use of evidence in pharmacovigilance: case reports as the reference source for drug withdrawals. Eur J Clin Pharmacol. 2001;57(1):89–91.
76. Juni P, Nartey L, Reichenbach S, Sterchi R, Dieppe PA, Egger M. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. Lancet. 2004;364(9450):2021–9.
77. Aronson JK, Derry S, Loke YK. Adverse drug reactions: keeping up to date. Fundam Clin Pharmacol. 2002;16:49–56.
78. Benson K, Hartz AJ. A comparison of observational studies and randomised, controlled trials. NEJM. 2000;342(25):1878–86.
79. Golder S, Loke YK, Bland M. Meta-analyses of adverse effects data derived from randomized controlled trials as compared to observational studies: methodological overview. PLoS Med. 2011;8(5):1–13.
80. Concato J, Shah MPH, Horwitz RI. Randomized, controlled trials, observational studies and the hierarchy of research designs. NEJM. 2000;342(25):1887–92.