Topoi (2014) 33:295–312
DOI 10.1007/s11245-013-9194-7

Hunting Side Effects and Explaining Them: Should We Reverse Evidence Hierarchies Upside Down?

Barbara Osimani

Published online: 18 October 2013

Abstract Philosophical discussions have critically analysed the methodological pitfalls and epistemological implications of evidence assessment in medicine, however they have mainly focused on evidence of treatment efficacy. Most of this work is devoted to statistical methods of causal inference with a special attention to the privileged role assigned to randomized controlled trials (RCTs) in evidence based medicine. Regardless of whether the RCT's privilege holds for efficacy assessment, it is nevertheless important to make a distinction between causal inference of intended and unintended effects, in that the unknowns at stake are heterogonous in the two contexts. However, although "lower level" evidence is increasingly acknowledged to be a valid source of information contributory to assessing the risk profile of medications on theoretical or empirical grounds, current practices have difficulty in assigning a precise epistemic status to this kind of evidence because they are more or less implicitly parasitic on the (statistical) methods developed to test drug efficacy. My thesis is that (1) "lower level" evidence is justified on distinct grounds and at different conditions depending on the different epistemologies which one wishes to endorse, in that each impose different constraints on the methods we adopt to collect and evaluate evidence; (2) such constraints ought to be understood to be different in the case of evidence for risk versus benefit assessment for a series of reasons which I will illustrate on the basis of the recent debate on the causal association between acetaminophen (a.k.a. paracetamol) and asthma.

B. Osimani (✉)
School of Pharmacology, University of Camerino,
Piazza dei Costanti, 62032 Camerino, MC, Italy
e-mail: barbaraosimani@gmail.com

Keywords Evidence hierarchies · Randomized controlled trials · Clinical methodology · Causal inference · Scientific inference · Induction · Abduction · Hypothetico-deductive method · Precautionary principle · Acetaminophen · Paracetamol

1 Introduction

Philosophers of science and practitioners have been debating for years about the superiority of randomized controlled trials (RCTs) over other forms of empirical studies in order to assess the efficacy of treatments. This debate has been made more dramatic by the diffusion and acknowledgement of so called evidence hierarchies which put RCTs at the top of evidence rankings.

Evidence hierarchies are intended as a decision heuristics for professionals overwhelmed by data of heterogeneous quality and pressed by time constraints (Howick et al. 2011). Nevertheless, they have been criticized for uselessly (and mistakenly) censoring a considerable quantity of potentially informative data, and at the same time for failing to recognize the intrinsic limitations of RCTs (Worral 2007a, b; Cartwright 2007). Regardless of whether these considerations hold for efficacy assessment, it is nevertheless important to make a distinction between causal inference of intended (treatment efficacy) and unintended effects (adverse reactions), in that the epistemic framework is heterogeneous in the two contexts.

Indeed, both philosophers as well as epidemiologists and health scientists have acknowledged the role of so called "lower level" evidence as a valid source of information contributory to assessing the risk profile of medications (Aronson and Hauben 2006; Howick et al. 2009;

Vandenbroucke 2006). However, they have difficulty in assigning a precise epistemic status to this kind of evidence and in providing a coherent justification for the sorts of exemptions they propose. In general, the quality of evidence coming from other sources than RCTs is considered high to the extent that possible known or unknown confounders can be safely excluded. Vandenbroucke (2008) adds to this scheme the distinction between the context of evaluation and the context of discovery as a possible explanation for the asymmetry between evidence standards for benefit versus risk assessment. On this basis he also proposes to reverse evidence hierarchies when risks rather than benefits of medical treatment are under scrutiny. (Vandenbroucke 2008).

This is only part of the issue however; in that different epistemological stances (hypothetico-deductive, inductive, abductive) provide different rationales for justifying evidence of diverse kinds and, because of this, may be more or less adequate to the purpose of risk assessment. My thesis is that, provided that knowledge about the drug risks comes from different sources and grows cumulatively in the course of time, inductive and abductive methods of causal assessment are better suited to the purpose than deductive ones. Furthermore, drug decisions are tied to the risk–benefit balance: according to the precautionary principle, the more severe is the new detected harm with respect to the expected benefit, the lower can be the probability of causal association between the harm and the suspected drug, in order to allow for safety measures (Osimani 2007, 2013a, b; Osimani et al. 2011). Thus, not only the hypothesis of causal connection need not be certain before triggering preventive measures, but, on the contrary, these measures must be enacted as soon as the hypothesis is sufficiently strong with respect to the risk–benefit balance. The conclusion is that the methodology for risk assessment should be reconsidered in light of these epistemological considerations and of their ethical implications.

The paper is structured as follows: Sect. 2 presents the philosophical debate on evidence hierarchies and their emphasis on randomized controlled trials (RCTs) as a privileged form of evidence of treatment efficacy (benefit assessment). Section 3 introduces Vandenbroucke's proposal to reverse evidence hierarchies when the evaluation of harmful effects is at stake (as opposed to benefit evaluation). I will then introduce the distinction between deductive and inductive/abductive approaches to scientific inference and connect them to other proposals of amendments to evidence hierarchies. Section 5 illustrates the rationales behind each kind of proposal by displaying the relationship between the underlying epistemologies and the principle of total evidence. Section 6 brings my argumentation to its conclusion by providing three distinctive reasons for preferring inductive/abductive ("vouchers") to deductive methods ("clinchers") of scientific inference when dealing with causal assessment of harm. Section 7 is devoted to the recent debate on the causal association between acetaminophen and asthma as a case in point.

2 The Debate on RCTs as a Privileged Form of Evidence in Medical Research

In evidence hierarchies for causal assessment RCTs are preceded by meta-analysis, of RCTs, only and are regarded as as the strongest evidence for therapeutic effectiveness, surpassing any kind of observational study, not to mention other sorts of evidence such as case reports (medical literature) or knowledge coming from basic science. More specifically, randomized controlled studies are followed by comparative studies which are not randomized (e.g. cohort or case–control studies), and these are followed by reasoning about patophysiologic mechanisms underlying the observed outcome. Expert judgment is regarded as the weakest form of evidence and put at the bottom of the hierarchy (see Howick 2011, for a recent philosophical overview).

The rationale for this ranking is provided by methodological-foundational considerations mainly developed by (frequentist) statisticians (but see Teira 2011, and Chalmers 2005, for a critical analysis of this issue). Within this perspective, randomization is supposed to play two roles: 1) repeated randomization of the treatment among the subjects in the sample, allows the study to approach in the limit (in the long run) the true mean difference between treated and untreated sample population (see Basu 1980, and Teira 2011; 2) randomization allows to experimentally isolate the cause under investigation from other prognostic factors (confounders). To this aim, the experimental machinery includes the following devices:

1. control (partition of the sample into treatment and control group/s)
2. intervention (treatment administration by the experimenter), and
3. double-blinding and placebo (concealment of treatment allocation from subjects and researchers).

Control allows for the comparison among treated and non-treated groups, thus allowing to infer causal efficacy by observing a difference in the two groups, provided that they are balanced in terms of prognostic factors; intervention severs the link between the treatment and the (known and unknown) reasons which motivate its use in real-life settings (self-selection bias); allocation concealment avoids allocation bias which might be caused by conscious or unconscious experimenters' expectations and interests; furthermore it minimizes the influence of these factors on the performance of the trial in terms of patient care and therapeutic support; and allocation concealment from subjects allows to minimize the influence on the experimental outcome caused by their expectations and attitudes toward the treatment as well as by the related behaviour. The specific role of randomization within this picture is to guarantee the baseline balance of treatment and control group(s) with reference to relevant prognostic factors (alternative/additional causes of the outcome of interest).

However, according to their critics, randomization is neither necessary nor sufficient to ensure that the observed experimental outcome can provide sound evidence about the treatment effect. The toughest attack has been launched by Howson and Urbach (2006) within a foundational comparison of frequentist versus Bayesian methods of statistical inference, and has been expanded on by John Worral in a series of papers, which specifically address the epistemic merits of RCTs (Worral 2007a, b, 2010). Worral's critique can be summarized in the following eight points:

1. Clinical researchers never randomize forever, so RTCs do not reflect the "limiting average" of the means differences between treatment and control group;
2. moreover there is "no sense in which we can ever know how close a particular RCT is to yielding this 'limiting average" (2007: 15);
3. Repeated randomization is, epistemically speaking, impossible: "If a particular patient in the study receives, say, the 'active drug' on the first round, then since this is expected to have some effect on his or her condition, the second randomization would not be rigorously a repetition of the fist. The second trial population, though consisting of the same individuals, would, in a possibly epistemically significant sense, not be the same population as took part in the initial trial" (2007: 22):
4. allowing sufficient "wash out" times between the rounds does not represent a perfect warrant against "contamination";
5. furthermore, repeated randomization is practically and ethically unfeasible;
6. randomization is only a means to the end of balancing the experimental groups and this aim can be reached also through other tools such as deliberate matching and "haphazard" allocation;
7. strictly speaking, it is not randomization but rather masking treatment allocation, which wards off bias due to experimenters' and subjects' interests and expectations (allocation and self-selection bias);
8. some meta-analyses comparing the results of RCTs and observational studies infer from the systematic overestimation of effects in the latter that they are less reliable; but judging the comparative reliability of observational versus randomized studies by taking the latter as the gold standard amounts to a *petitio principii*.

Whereas points 1–5 represent indeed a formidable objection to the application of frequentist statistics to clinical trials, and point 8 needs no defence, still the role of randomization in helping to ward off various forms of bias cannot be entirely trivialized (points 6 and 7). A pragmatic defence of RCTs, which nevertheless acknowledges their methodological limitations can be found in La Caze (2009) and Teira (2011). La Caze emphasizes the distinctive roles played by intervention and by randomization. While it is intervention that actually severs the causal link between the decision to take the drug and all antecedents leading to this choice in real-life context (i.e. all confounding related to self-selection bias); randomization contributes nevertheless to the internal validity of the experiment by increasing the chance that comparison groups are balanced in terms of prognostic factors. Quoting Suppes (1982) and Lindley (1982), La Caze (2013) adds that randomization can also be defended on Bayesian grounds as a way to simplify the computation of the prior probability function concerning the belief itself that relevant prognostic factors are indeed balanced in the experimental groups.[1]

While La Caze provides reasons to privilege RCTs over non-randomized studies in terms of reliability and computational tractability, Teira (2011) analyzes 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 and in particular, he points to the concrete possibility that randomization may yield, just by chance, an unbalanced distribution of the prognostic factors. However, he goes on, randomization "is still a warrant that the allocation was not done on purpose with a view to promoting somebody's interests". Thus, Teira 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 (I will come back to this point at the end of the paper).

More cogently, the weakness of RCTs is evident when considering the issue of external validity. This point has been

[1] However, Suppes (1982) rather defends randomization as a way to balance the two groups in causal inference, whereas he ascribes a computational advantage to randomization within sample-to-population inference.

analytically treated by Cartwright (2007), who enumerates the assumptions which should be met in order to export the claim of efficacy from the sample to the target population. These can be simplified in the requirement that at least one causally homogeneous subgroup in the target population has the same causal structure and probability measures of at least one causally homogeneous subpopulation in the experimental sample. Thus the evidence provided even by an ideal (i.e. perfectly internally valid) RCT can be extended to the target population only with great caution. The conclusion is that not only is randomization neither necessary nor sufficient to the overall end of causal inference for medical treatments, but that it is also not recommended for most practical purposes it is supposed to pay service to, and that its interpretation and use may need more information than it delivers (see also Cartwright 2010).[2]

On the other hand, traditional advocates of RCTs and evidence hierarchies are gradually recognizing that the virtues of RCTs cannot secure their privileged position in causal inference and efficacy assessment without any further specifications. Thus they recommend that randomization be not considered as the only criterion to evaluate evidence quality: other characteristics such as effect magnitude and consistency across studies, dose–response gradient, as well as publication bias should be taken into account as well. Also the problems of external validity and extrapolation are regarded as particularly serious for implementing results of RCTs on target populations (see for instance Howick 2011). This awareness has led to the development of new guidelines for ranking evidence (e.g. the GRADE System: Guyatt et al. 2008, 2011; see also the 2011 Oxford CEBM Levels of Evidence: Howick et al. 2011).

Now, this debate on RCTs and evidence hierarchies has failed so far to clearly distinguish the context of efficacy versus risk assessment and treated them as one and the same problem. Undeniably, awareness of this distinction is gradually growing, and "lower level" evidence is increasingly acknowledged to be a valid source of information contributory to assessing the risk profile of medications on theoretical (Aronson and Hauben 2006; Howick et al. 2009) or empirical grounds (Papanikolaou et al. 2006; Benson and Hartz 2000; Golder et al. 2011; Concato et al. 2000). Indeed, in their comparative analysis of RCTs and observational studies, Papanikolaou et al. (2006) for instance assert that "it may be *unfair* to invoke bias and confounding to discredit observational studies as a source of evidence on *harms*" (p. 640, my emphasis). But nobody until now has explained why this should be so. In my opinion there are two answers to this question. One is related to substantive issues concerning such problems as causal interaction, modularity and external validity as well as extrapolation when evaluating the reliability of experimental or observational methods; the other relates to epistemic issues concerning the constraints we put on evidence and to foundational issues in scientific inference. I will address in this paper the latter dimension and argue that:

1. different epistemologies may justify "lower level" evidence on different grounds;
2. in the case of risk detection and assessment non-deductive epistemologies are better suited to the purpose.

I will start my argument by presenting Vandenbroucke's proposal to reverse evidence hierarchies and consider his argumentative points in light of the epistemological paradigms they more or less implicitly are based on.

3 "Epistemic Asymmetry" of Risk Versus Efficacy Assessment

In the last decade a series of papers written mainly by epidemiologists has developed the view that evidence for harm and for efficacy should be evaluated according to different criteria (Vandenbroucke and Psaty 2008; Vandenbroucke 2004, 2006, 2007, 2008; Psaty and Vandenbroucke 2008; Papanikolaou et al. 2006; Stricker and Psaty 2004; Agency for Healthcare Research and Quality 2007). These studies underline the different value of randomized controlled experiments and observational studies (included case reports) for benefit and risk assessment. Some have also proposed a reversal of the hierarchy for risk detection with respect to benefit assessment (Vandenbroucke 2008), and have advocated a sort of methodological pluralism according to which study designs are evaluated in relation to the research goal and the incognita under investigation. As a consequence, no epistemic advantage of randomized versus observational studies is claimed in principle, but rather their evidential strength is evaluated with respect to whether they are used to evaluate

[2] Papineau (1994) insists on the different roles played by random sampling as a means to achieve sample representativeness (i.e. external validity), and experimental randomization as a means to avoid self-selection bias and to deal with unknown confounders in causal inference (i.e. internal validity). He defends the latter, while acknowledging the epistemic paradoxes affecting the former. It should be however noted that patient recruitment is far from complying with the principles of random sampling in a strict sense. Furthermore, as also Urbach (1994) in his reply to Papinau underlines: "population probabilities are, in my opinion, not easy to conceptualize when we are dealing with the responses of types of patient to medical interventions" (p. 714). Which is indeed the issue analyzed in detail by Cartwright (2007). It is worth noting that, the problem of external validity affects observational studies and studies of mechanisms as well (Howick et al. 2013), but in the case of RCTs this drawback has more detrimental implications to the extent that they are presented as the "*gold standard*".

Table 1 Evidence hierarchy reversal for benefit versus risk assessment (Vandenbroucke 2008: 5)

Hierarchy of study designs for intended effects of therapy	Hierarchy of study designs for discovery and explanation
(1) Randomised controlled trials	(1) Anecdotal: case report and series, findings in data, literature
(2) Prospective follow-up studies	(2) Case–control studies
(3) Retrospective follow-up studies	(3) Retrospective follow-up studies
(4) Case–control studies	(4) Prospective follow-up studies
(5) Anecdotal: case report and series	(5) Randomised controlled trials

the claimed therapeutic benefit or to assess/discover unintended effects. In particular, Vandenbroucke (2008) endorses the idea that evidence hierarchies should be reversed when the problem is not to test an intended effect but rather to discover an unintended one (Table 1).

Vandenbroouke presents several arguments in support of such a proposal. One is strictly methodological, the others touch epistemological questions related to scientific inference and evidence evaluation.

The first point, which he admits not to be new in the pharmacoepidemiological literature (Jick 1977; Jick and Vessey 1978; Miettinen 1983), but which deserves further attention among both scientists and practitioners, is methodological, and concerns the idea that selection bias is less likely to affect observational studies with respect to adverse reactions.[3] This is because unintended effects, qua unintended, are not known in advance, and thus also not known by the drug prescriber, who cannot take them into consideration and thus bias treatment allocation. There is a *continuum* of course, where hepatic reactions are predictable side effects especially for specific subgroups such as elderly patients, whereas immune reactions are mostly unpredictable and difficult to foresee. Because of this, observational studies concerning adverse reactions will not suffer from confounders in the same way as observational studies for intended effects do (Vandenbroucke and Psaty 2008; Vandenbroucke 2004, 2006, 2007, 2008; Psaty and Vandenbroucke 2008). Even if the doctor knows whom s/he is prescribing the treatment to, treatment allocation is masked with respect to unintended effects, given that s/he does not know them: "As a mirror image for adverse effects research, the doctor *knows that he is prescribing a drug to a particular patient, but he might not know the risk* that this patient has of developing a particular adverse effect. […] This achieves the same aim of breaking the link between prescribing and prognosis" (Vandenbroucke 2004: 1728). Ignorance of the unintended effects guarantees the same unbiasedness of ignorance about whom the treatment is actually administered. Thus "for many problems in genetics, infectious disease outbreaks, or *for adverse effects of drugs, no further evidence may be needed*" (Vandenbroucke 2008: 6, my emphasis). The focus of this argumentative point is randomization as a method to avoid selection bias and the thesis is that, provided that this bias can be excluded in the case of unintended effects, randomization is not necessary in this respect.

The second point advanced by Vandenbrouke in defence of his hierarchy reversal draws on epistemological considerations and is based on his distinction between the context of discovery and the context of evaluation in harm detection and assessment. (Vandenbroucke 2008: 1–7). According to Vandenbroucke, discovery is focused on explanation and hypothesis generation; evaluation instead, on hypothesis testing or confirmation, and thus it may be hold that research methods differ in the opportunities they offer with respect to either of these goals:

> "For discoveries, the original case reports, lab observations, data analysis, or juxtaposition in literature may be so convincing that they stand by themselves, either because of the magnitude of the effect or because the new explanation suddenly and convincingly makes the new finding fall into place with previous unexplained data or previous ideas". (Vandenbroucke 2008: 6).

To the extent that side effects are being discovered rather than tested, "lower level" evidence may be totally satisfactorily. Indeed, case reports remain the most sensitive (and sometimes the only available tool) for discovering side effects, and, far from constituting a second-best choice, case series (and single case reports) as well as findings in data and literature are a privileged tool for risk detection (see also Stricker and Psaty 2004 and Glasziou et al. 2007).

Vandenbroucke's third point follows from the second in that he formalizes the contrast between the context of evaluation and the context of discovery in terms of different priors assigned to hypotheses of benefits and of adverse reactions. When an intended effect is tested, prior odds are quite high (there is a 1:1 odds that the therapy will be at least as efficacious as the standard treatment). This also reflects the requirement of equipoise for undertaking such trials, i.e. that the standard therapy and the treatment being tested have the same expected net utility (see Freedman 1987, but also Gifford 1995, 2007a, b for important specifications on this topic). Instead, Vandebroucke points out, when new ideas emerge from unexplained data or from the examination of existing

[3] This topic relates also to Teira's (2011) impartiality argument (see Sect. 2 and 6 in this paper).

literature, priors are quite low and they are often disconfirmed by subsequent studies. Vandenbroucke deduces two consequences from this state of affairs. The first one is that it is the higher priors which make the results more robust, not the method (Vandenbroucke 2008: 16–17). The second one is that the reason why we accept uncertain results for risks rather than for benefits is that evaluation and discovery studies are associated with different loss functions: evaluation is related to the approval of health technologies and is required to assure stakeholders about their efficacy and safety, whereas discovery is more related to the context of research for its own sake, which might explain why certain study designs are preferred to others in these contexts.

I will come back to these considerations in Sect. 6; for the time being, my main interest is to emphasize how Vandenbroucke's defence of his hierarchy reversal is grounded on different epistemic rationales, namely on criteria borrowed from the hypothetico-deductive methodology underlying statistical hypothesis testing, or on inductive and abductive reasoning.

Hypothetico-deductive methods such as classical hypothesis testing work with truth conditions and put severe constraints on the kind of evidence which can be admitted, or at all considered meaningful in assessing causal associations. Instead inductive and abductive methods work with fallible indicators and put different sorts of evidence together in order to infer the implications which derive from their joint support to the hypotheses under investigation. Roughly speaking, the former methods aim at hypothesis rejection or acceptance with no degree in-between, while the latter aim at a judgment on the plausibility of the hypothesis given the data, which can be possibly quantified in probabilistic terms. Thus the reasons for accepting "lower level evidence" differ in the two settings.

In Vandenbroucke's argumentation, a deductive approach is evident in the justification he provides for considering RCTs and observational studies on the same level: his point is that they have the same reliability when the causal effect is unknown, because this state of affairs amounts to avoiding selection bias (I will explain why this sort of argumentation is grounded on a deductive perspective in the next section); an inductive approach pops up when he draws on loss functions and priors in order to account for the intuition that anecdotal evidence may make an excellent service to the purpose of risk detection and assessment. In fact, another rationale for accepting "lower level" evidence as a valid support for hypotheses of harm lurks in Vandenbroucke's contribution. This is represented by the quotation cited above, where he says that original case reports, observational data or juxtaposition in literature might be sufficient because "the new explanation suddenly and convincingly makes the new finding fall into place with previous unexplained data or previous ideas" (Vandenbroucke 2008: 6). This sort of strategy can be considered to follow an "abductive" approach to scientific inference, in that it works by putting together different sorts of evidence and infer the implications which derive from their joint support to the hypothesis under investigation.

4 Deduction, Induction and Abduction in Pharmaceutical Harm Assessment

Vandenbroucke is not the only one to propose exemptions to the standard canon of evidence hierarchies. Other examples include even authors who are traditionally associated with the orthodox paradigm of evidence based medicine. I will present these proposals by tracing them back to the epistemological paradigms just mentioned. In analogy with the analysis of Vandenbroucke's argumentative points, I distinguish three main epistemological paradigms: (hypothetico-) deductive (or, also said, falsificationist), inductive, and abductive methods of scientific inference. For our purposes however, the main distinction is between deductive and non-deductive approaches—"clinchers" and "vouchers" in Cartwright's terms (Cartwright 2007)—thus, I will draw the line between these two categories.

4.1 Clinchers

4.1.1 Hypothetico-Deductive Approach: Hypothesis Testing

The aim of hypothesis testing is to provide a means to reject hypotheses on the basis of statistical evidence. The logical ground for this procedure is provided by the inference rule of *modus tollens*: H entails E, your evidence contradicts E, hence you may infer that H is not the case.

$$\begin{array}{l} \mathrm{H} \rightarrow \mathrm{E} \\ \dfrac{\neg\mathrm{E}}{\neg\mathrm{H}}. \end{array}$$

In the case of hypothesis testing E is represented by the difference among treated and non treated groups. If you observe no significant difference, than you reject the causal hypothesis. The rationale underpinning this sort of method is that the difference between the comparison groups in the trial is due to the contribution of the investigated factor, and only to it. Consequently, the more likely a method is to be able to exclude confounders (i.e. additional contributing factors to the observed result), the more reliable is the inference we base on it, and the higher the method is ranked in the hierarchy. Indeed the main *raison d'être* of the very idea of ranking evidence is to provide an a priori evaluation criterion, based on the exclusion of confounders. As a corollary, case reports and observational data are considered sufficient evidence for causal claims *to the extent that* possible confounders can be confidently excluded.

Glasziou et al. (2007) for instance consider cases where the relation between treatment and effect is so dramatic that bias and confounding can be safely excluded even in the absence of randomization: these are represented as phenomena of sudden and drastic change in the clinical/epidemiological pattern and are formalized in terms of signal to noise ratio. Howick et al. (2009) relax the requirement of dramatic effect and reduce it to the desideratum that the effect size be greater than the combined effect of plausible confounders. Along the same lines it is however specified that observational studies must demonstrate larger effects than randomized trials because of their greater risk of confounding from selection bias (since the allocation to treatment groups is neither randomized nor concealed) and performance bias (since the participants and caregivers are not blinded) (p. 188). Even though he comes to an opposite conclusion, also Vandenbroucke's point concerning the equal trustworthiness of RCTs and observational studies for unintended and unexpected consequences of interventions is based on similar considerations. As a matter of fact, he ascribes RCTs and observational studies the same reliability on grounds that ignorance about the unexpected consequences of an interventions *achieves the same lack of bias* obtained through blinding (i.e. ignorance about whom will receive the treatment).

4.2 Vouchers

4.2.1 Inductive (-Bayesian) Approach

The discovery/evaluation distinction proposed by Vandenbroucke (2008) is cast in Bayesian terms, in that it explores the epistemic asymmetry between benefit and risk assessment in terms of different priors associated with intended versus unintended outcomes. As a matter of fact, the distinctive point between the inductive-bayesian framework and classical hypothesis testing, is that in the latter, hypotheses are formulated and then tested for rejection/acceptance. Instead, in the former hypotheses are assigned a probability, on the basis of available knowledge/data, and this is then updated in light of new evidence. Also, evidence is interpreted in light of all possible alternative hypotheses. Probability measures specify the degree of support enjoyed by hypotheses.

The general principle of induction is that it cannot deliver you a sure-fire guarantee about the conclusion of your inference. However, it may help you assess the plausibility of a given hypothesis on the basis of inconclusive but relevant evidence. For historical reasons, Bayesian approaches to trial design and postmarketing surveillance have not received much attention by the regulator and by the medical community. However the new guidelines for pharmacovigilance (EMA and HMA 2012), put a special emphasis on joint efforts for what can be considered an information based (rather than power-based) approach to pharmaceutical risk assessment, and encourage the integration of information coming from different sources (spontaneous case reports, literature, data-mining, pharmacoepidemiological studies, post-marketing trials, drug utilization studies, non-clinical studies, late-breaking information; see also Herxheimer 2012).

4.2.2 Abductive Approach

This strategy rather rests on an approach to scientific inference which instead of experimentally isolating the causal factor under investigation, works by putting together different pieces of evidential facts and then inferring the implication of their joint occurrence. Rather than filtering evidence by ranking it, this approach aims to accommodate all data in a unifying picture. It is more or less knowingly advocated by different authors in relation to the detection and causal assessment of harms. Aronson and Hauben (2006) for instance put forward that "In some cases other types of evidence may be more useful than a randomised controlled trial. And *combining randomised trials with observational studies and case series can sometimes yield information that is not available from randomised trials alone*" (my emphasis). This idea is also at the basis of the recent proposal by Howick et al. (2009), Russo and Williamson 2011, and Stegenga (2011) to integrate evidence hierarchies with Bradford-Hill criteria for causal inference. As a matter of fact, Bradford-Hill criteria are not meant as truth conditions for causality but rather as imperfect indicators which jointly support the hypothesis of causation. Hence, Howick's and Stegenga's proposals move away from the idea of obtaining perfect information about causality in a one-shot test and go in the direction of abductive/inductive reasoning.

It is worth recalling at this point that Bradford Hill criteria are meant as an alternative approach to hypothesis testing for causal assessment. At page 299 of his most cited paper, Sir Bradford Hill specifies the rationale behind his nine guidelines:

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

Table 2 Epistemological paradigms and related rationales for justifying "lower level" evidence

Epistemology	Method	Main assumptions	Justification of "lower level" evidence
Hypothetico-deductive (statistical mode)	Hypothesis testing: likelihood of evidence if H_0 = true (p-value)	Investigated factor is isolated by balancing the experimental groups as to all other prognostic factors	Only if alternative explanations for the observed result (confounders) can be safely excluded, or treatment effect swamps them by a statistically significant amount
Abduction	Connection of data in light of explanatory hypothesis	Account for as much evidence as possible	Explanatory power of hypothesis in light of data
Inductive-Bayesian	Bayes theorem	Principle of total evidence—coherence	Probability of hypothesis given likelihood function and prior

> *than cause and effect*?" (my emphasis. Thus, Bradford Hill both refers to explanatory power and likelihood as reliable grounds to justify causal judgments, thereby adopting, at least implicitly, an alternative approach to causality as that implied by RCTs. And immediately after that he adds: "No formal tests of significance can answer those questions. Such tests can, and should, remind us of the effects that the play of chance can create, and they will instruct us on the likely magnitude of those effects. Beyond that, they contribute nothing to the proof of our hypothesis", which is not the only point in his paper where Bradford Hills presents is criteria as opposed to tests of significance.

Table 2 illustrates the epistemological rationales for justifying "lower level" evidence according to hypothetico-deductive, abductive and inductive approaches to scientific inference.

Whereas the acceptability of lower level evidence for harm assessment is increasingly acknowledged by scholars of different epistemic stances, the reasons provided for the exemption from standard evidence rankings are different.

In classical hypothesis-testing, The result is expressed as the probability (p value) that the experiment delivers the observed result—or more "extreme" results—if the treatment makes no difference (so called null Hypothesis: H_0). For the result to be at all meaningful, it is essential that the observed difference among groups is due to the treatment and only to it. Which in turn explains the insistence on the exclusion of confounders. Abductive methods instead, work with imperfect indicators, which need to be connected by means of an explanatory hypothesis: thus not only can any relevant observation be used for justifying the hypothesis under consideration, but it also must be so (the greater the proportion in a data-set which a hypothesis accounts for, the greater its explanatory power). Inductive approaches differ from classical hypothesis testing, in that hypotheses are neither refuted nor accepted, but instead are associated with a probability which is updated in light of data, following Bayes theorem (on its turn grounded on the calculus of probability). Also in this case, any piece of possibly relevant evidence can provide a certain amount of support to the entertained hypothesis. The main constraint is provided by the requirement of coherence.

The essential distinction between clinchers and vouchers is that whereas the former put strict desiderata on evidence for it to allow meaningful inferences, the latter are guided by the idea that all relevant evidence—or as much data as possible in the case of abduction—should be taken into account in order for the inferential procedure to be valid. This is also called the principle of total evidence.

5 Total Versus Best Evidence

Keynes (1921) traces back the origin of the principle of total evidence to Bernoulli's maxim that "in reckoning a probability, we must take into account all the information which we have" (Carnap 1947: 138, footnote 10; quoting Keynes 1921: 313). The principle of total evidence has been a topic of hot debate among philosophers such as Hempel, Carnap, Ayer, Braithwaite, and Kneale among others. An outline of the history of the debate is out of the scope of this paper. For the present purpose it is sufficient to point out its relation with the non-monotonic character of inductive inferences.

Nonmonotonicity means that conclusions of inductive inferences (either quantitative ones, such as in probabilistic approaches, or qualitative, such as in adaptive logics) are contingent and may be invalidated by additional information (Kyburg and Teng 2001; Meheus 2011). As a matter of fact, induction can be characterized as an inference where the evidence does not entail the hypothesis, but only more or less strongly supports/undermines it (Ayer 1956, 1957). Inconclusive evidence is used to assess the plausibility of a hypothesis and to possibly quantify it in a probabilistic fashion, so that, for instance $P(H|E) = .9$; but there may always be additional information F, which may lower this support, so that, for instance, $P(H|E\&F) = .2$. This means that "acquired support" may get lost if additional information undermines it. Let's illustrate this phenomenon

Table 3 Non-monotonicity in inductive inference

Deductive inference: conclusive evidence	Inductive inference: inconclusive evidence
Modus ponens $E \rightarrow H$ $\frac{E}{H}.$ No other additional evidence can change the conclusion. If, in addition to E, you come to know F, you always have H as a conclusion: $E \rightarrow H$ $\frac{E, F, \ldots}{H}.$	When E represents non-conclusive evidence for H, there may always be the possibility that $P(H \mid E) > P(H),$ and that additional evidence F might reverse this inequality thus leading to the following result: $P(H \mid E) < P(H \mid E, F)$
The same is valid for *modus tollens*: $H \rightarrow E$ $\frac{\neg E.}{\neg H}$ No additional evidence would change this conclusion	The bearing of this phenomenon is most evident when comparing the strength of support provided by the evidence to the hypothesis H and its complement (¬H). So that you may have: $P(H \mid E) > P(\neg H \mid E)$ And, after learning F: $P(H \mid E, F) < P(\neg H \mid E, F).$

with an example. A doctor thinks that a patient is celiac, because all his/her available evidence E (adverse reactions to certain foods, iron deficiency, a series of additional symptomatic phenomena) points to this diagnosis, however he then prescribes a series of serum tests and they all result negative (evidence F). Then the strong support to the diagnosis of celiac disease provided by E is "corroded" by the negative evidence F and the doctor needs to look for a hypothesis which accounts for both E and F: for instance a simple food intolerance.

Table 3 provides a formal comparison of deductive and inductive inference with reference to the problem of non-monotonicity.

The example in the table shows the case where a given hypothesis H is favoured over its complementary (¬H) after learning E, and then it is disfavoured after learning also F ("defeating evidence").

Statistical hypothesis-testing is a kind of approach which admittedly follows a Popperian hypothetico-deductive method of scientific enquiry. And being this paradigm inherently deductive, it does not feel urged to address the issue of non-monotonicity. Once you have conclusive evidence E rejecting hypothesis H, any other piece of evidence becomes irrelevant. Thus the closer the evidence gets to this deductive ideal, the better: best evidence means evidence which maximises internal validity.

Indeed 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 (Evidence Based Medicine Working Group 1992; Sackett et al. 1996; Straus and McAlister 2000). However by putting their emphasis on ranking and on internal validity they endorse a lexicographic rule of implementation of evidence hierarchies. For instance, if you have evidence E from a cohort study which results in a significant difference between exposed and non-exposed group, and you obtain evidence F from an RCT which fails to reject the null hypothesis that the difference between treated and non-treated groups is non-significant, then the non-difference hypothesis holds (F discards E). This is because F is supposed to represent the difference produced by the treatment alone (which equals 0 in this case), whereas E does not have this guarantee and may result from other prognostic factors. While evidence hierarchies have also been given a heuristic interpretation (see: Howick et al. 2011), this does not change the fact that the rules are epistemically grounded on internal validity maximization.

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 case study below as an example for this attitude).

Even if more sophisticated versions have been developed which are at pains to distinguish between different hierarchies depending on different evaluation goals—see for instance the CEBM[4] levels of evidence subdivided in therapy, prognosis, diagnosis, and economic analysis—still, efficacy and harm assessment are coalesced in one and the same column: therapy-prevention-etiology-harm, where meta-analyses of RCTs, followed by single RCTs, are at the top of the ranking. Similarly, Guyatt and colleagues (2011) admit the difficulties inherent in the evaluation of evidence for harm, but propose a framework (the GRADE System) where its quality is assessed with the same criteria proposed for efficacy evaluation. Particularly, evidence for harm coming, say, from observational studies is given lower weight than evidence for efficacy coming, say, from RCTs, thus biasing the overall risk–benefit assessment in favor of the drug.

More generally, the very idea of ranking or up- and downgrading evidence on the basis of its internal validity is at the opposite side of a unifying approach which aims to account for all the evidence at disposal. In fact, non-deductive approaches must take into account all available evidence, because no matter how much a piece of evidence supports a given hypothesis, the possibility of defeating evidence can never be excluded.

[4] Howick et al. 2011; http://www.cebm.net/mod_product/design/files/CEBM-Levels-of-Evidence-Introduction-2.1.pdf.

A straightforward consequence is that reversing hierarchies may not represent the real solution to the problem of imperfect risk information. Rather, one should consider the specific epistemic structure of the problem at hand and then consider whether clinching or vouching methods should be preferred.

6 Should We Reverse Evidence Hierarchies?

Considering the above analysis of evidence evaluation and epistemic criteria underlying deductive versus inductive/abductive approaches to scientific inference, it becomes clear that the issue is not whether hierarchies should be reversed or not, but rather what kind of approach best serves the purpose of causal assessment with respect to harms. I can find at least four reasons why clinchers should be preferred to vouchers when assessing harm.

6.1 Integration of Prior Knowledge and Observation

Frequentist statistics does not allow to incorporate priors in hypothesis evaluation. This is a particularly detrimental drawback in the case of harm assessment considering that much knowledge of the drug behavior may be inferred analogically from same-class molecules or similar entities. Also, theoretical awareness about the drug unknowns should be taken into account when evaluating risk signals. For instance, most compounds are racemates, meaning that they have an equal amounts of left- and right-handed enantiomers of a chiral molecule, and generally only the effects of one of the two enantiomers are sufficiently known through the approval procedure, while the effects of the other are ignored. Culpable negligence of this state of affairs is at the origin of the Thalidomide tragedy for instance. Finally, historical knowledge about drugs harmfulness in general cannot be neglected in the process of causal assessment. In fact, the acceptability of anecdotal evidence or of uncontrolled studies for assessing risk has to do with a high prior about the *general capacity* of the drug to bring about side-effects: Whereas there is *total ignorance as to any specific side effect* which might be possibly caused by the drug, still there is almost certainty about the fact that the drug will indeed cause side-effects beyond the ones already detected in the pre-marketing phase. This high prior derives from historical knowledge and past experience with pharmaceutical products: it may be more or less precise depending on the novelty of the molecular entity and more or less high depending on the risk profile of better known analogous drugs or drug classes. Anyway, when combined with the high prior belief that there are unknown risks yet to be detected, “lower level evidence” may constitute a sufficient basis for action in proportion to the magnitude of the detected risk and the plausibility of the causal association.

Indeed the high default prior for an undefined risk also explains the rationale behind the introduction of the precautionary principle in the pharmaceutical domain and is also strongly reflected in the regulation which introduced the notion of “development risk” (or “potential risk”), i.e. the unknown latent risk unavoidably associated with the drug, as well as the pharmacosurveillance system. The drug is approved “with reservation” (approval can be withdrawn at any moment on the ground of newly discovered adverse reactions) and it is constantly monitored precisely because of the high prior associated with the possibility of it causing other side-effects beyond the ones discovered in phase I-III of the approval procedure. Thus the acceptability of non-experimental evidence is not due to the fact that stakes are lower, but on the contrary, just because these are high, evidence choice is allowed to be highly flexible in order to allow any data to play a role in early risk detection and prevention.

6.2 Cumulative Causal Learning and Categorical Versus Probabilistic Causal Assessment

From the time a risk is not known, to the moment in which it is incontrovertibly proven to be causally associated with the drug, there is a period of evidence accumulation which constitutes a state of partial and imperfect (but continuously increasing) knowledge. In this period it cannot be claimed that there is a causal link between the drug and the detected risk; but neither can we behave as if we knew nothing about it. Still, the latter attitude is precisely the only possible policy allowed by an epistemology grounded on hypothesis rejection. Moreover, following the precautionary principle, which has been developed precisely by taking into account these considerations, you are not supposed to wait for the causal connection between harm and suspected drug to be certain, before you take adequate countermeasures, but instead, you should act as soon as the probability of causal connection is high enough to recommend countermeasures because of a negative risk/benefit balance. This probability might be also very low, in case the risk magnitude is considerably big with respect to the expected benefit. The frequentist mode of summarizing statistical data, following which hypotheses may only be accepted or rejected, cannot be of any use to this purpose.

6.3 Impartiality

The issue of impartiality assumes in the case of benefit versus risk assessment opposite characteristics. Since benefit is intended and desired, but may be counterfeited for obvious commercial interests, the most natural way to deal with bogus products is to put the claim of efficacy to the test of strict trials (which is indeed what originated the success of randomized trials in the regulatory domain: Teira 2011). As for the risk, the situation is quite opposite: on the side of the industry, there is all interest in discounting the drug as a possible causal contributor to the side effects, thus the stricter are the standards for causal assessment, the easier it is for them to provide whitewashed drug profile.

Teira (2011) conceptualizes impartiality as a way to deal with uncertainty such that it cannot be exploited by some party's private interest. Well, 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 is exploited by the industry's private interest.[5] I think this is what Papanikolaou et al. (2006) have in mind when they say that "it may be unfair to invoke bias and confounding to discredit observational studies as a source of evidence on harms". By regimenting benefit and risk assessment with the same standards, we forget that in the case of risk, the question we want to answer is not whether the drug *really* causes it, but whether we can safely exclude that it does.

Yet, the established commonplace that causation can only be proved by higher level evidence, ends up with dismissing causal hypotheses unless supported by such evidence, and with disregarding the evidential force of other epistemic cues such as the likelihood of the total available evidence on the hypothesis of causation, or its explanatory power.[6] I will present the recent discussion on the debated causal association between acetaminophen and asthma as a case in point.

7 The Case of Acetaminophen and Asthma

The asthma epidemic, which started in 1960 is still an enigma for epidemiologists and immunologists alike. It is out of the scope of this paper to exhaustively present the puzzles raised by the change in prevalence and severity of this disease across Western countries (to which an enormous list of publications is devoted and for which there are specially devoted journals). I am rather interested in the current debate concerning the suspicions about acetaminophen (also known as "paracetamol") being a possible contributor to the inception and exacerbation of this disease, especially in paediatrics. In fact, a significant association between acetaminophen and increase in asthma incidence/severity has been established in observational studies (Henderson and Shaheen 2013; Holgate 2011; Farquhar et al. 2010). The question now is whether, given the available evidence, we should wait for an RCT to prove that this association is causal, or whether we should already recommend against its use among at risk people, especially children and pregnant women. McBride (2011) and Martinez-Gimeno and García-Marcos (2013) favours this latter position, whereas most commentators align with the EBM protocol and urge for RCT trials in order to settle down the issue. I will briefly present the case and then relate it to the preceding philosophical discussion.

7.1 How Suspicion Fell on Acetaminophen

Statistical data quantifying the change in asthma prevalence and severity that there has been in the United States and in other Western countries have struck the attention of health practitioners and epidemiologists since the early nineties: it is reported a 75 % increase among adults in the U.S. in the last 3 decades and a 160 % increase among children in the same period (Burr et al. 1989; Eneli et al. 2005; Ninan and Russell 1992; Mannino et al. 1998, 2002; Seaton et al. 1994; Eder et al. 2006; Subbarao et al. 2009). Provided that host susceptibility is unlikely to change so abruptly, epidemiologic research has focused on environmental factors (and more recently, on gene-environment interactions) that might be supposed to have changed as well, in the same or a compatible time-period: (1) increased exposure to outdoor and indoor pollutants; (2) decreased exposure to bacteria and childhood illnesses during infancy (the "hygiene hypothesis"); (3) increased obesity incidence and prevalence; (4) changes in diet and oxidant intake as well as physical activity (Platts-Mills et al. 2005); (5) cytokine imbalance as a reaction to environmental allergens in early childhood leading to lifelong T-helper type 2 (allergic) dominance over T-helper type 1 (nonallergic) reactions, thus increasing the risk for atopic disease (see Eneli et al. 2005; Seaton et al. 1994; Shaheen et al. 2000). However these have provided contrasting signals ranging from protecting to inducing asthma, sometimes depending on the exposure age (Martinez-Gimeno and García-Marcos 2013).

For instance, on the basis of a careful examination of all available evidence, Seaton et al. (1994) dismissed causes 1

[5] This is all the more true when risk data are already available from observational studies, which themselves need a sufficiently long time before they can deliver significant results (prospective designs) or can be started only when the drug has been used for a sufficiently extended period (retrospective designs).

[6] A noteworthy contribution in this respect is constituted by Russo and Williamson's effort to provide an epistemic approach to causality, which considers both evidence of statistical association and evidence of underlying mechanisms as jointly contributory and reciprocally complementary to providing evidence for causality (Russo and Williamson 2011). See also a further development of this line of thought in: Clarke et al. (2012 and forthcoming). Joffe (2011) provides a careful review of major biological mechanisms relevant for causal inference in epidemiological investigation.

and 5 as major candidates for inducing asthma epidemic because of a mismatch between the patterns of asthma increase and the increase of any of the incriminated indoor or outdoor pollutants/allergens in the environment. Because of this mismatch, it is implausible that they are responsible "alone or in combination, for the substantial rise in the prevalence of atopic disease" (172). Seaton et al. (1994) thus point their attention to a reduction of host resistance due to the "westernized" diet: reduced consumption of asthma-protective food such as fresh vegetables, fruits, which are important sources of antioxidants (e.g. vitamin C and β-carotene), as well as red meat and fresh fish, sources of ubiquinone, selenium and zinc, which are essential cofactors for antioxidant defence mechanisms. Indeed, whatever its diverse and complex etiology, asthma is characterized by airway inflammation, one important mechanism of which is the generation of oxygen free radicals. Thus, lack of nutritional antioxidants can be reasonably considered a relevant cause for asthma exacerbation. However the diet hypothesis is a very complex one to prove because diet is difficult to measure; particularly it is difficult to identify the combined and independent effects of the different nutrients (Eder et al. 2006). Also, the same element may have contrasting effects on the same outcome, such as for instance selenium which is an antioxidant but may also upregulate immune responses typical of allergic asthma. The hygiene hypothesis has been questioned also because of scarce consistency between the time trends of other allergic diseases (such as hay fever) and asthma (Platts-Mills et al. 2005). More generally, it is by now established that none of the environmental factors alone is able to explain the time-trend, and more attention should be devoted to their interactive effects on the incidence and severity of asthma (Eder et al. 2006).

Unlike the association between asthma and various environmental factors, its association with acetaminophen consumption seems to be clearer and consistently positive across studies. More interestingly, the time of acetaminophen introduction in clinical practice and its consumption trends seem to perfectly reflect asthma epidemic and therefore to provide a distinctive explanation for it.

The first clue indicating a possible relationship between acetaminophen and asthma has been indirectly provided by a study conducted in 1998 by Varner et al. (1998) in which they detected a precise correspondence between increase of asthma incidence and increased acetaminophen use as a substitute for aspirin (substitution which occurred once an association was recognized between aspirin and Reye's syndrome). The trend levelled off in the 1990s, i.e. at a time when acetaminophen had already become one of the most widespread analgesics. Varner et al's tentative explanation of this phenomenon was that asthma increase was due to aspirin avoidance, as aspirin may protect from asthma through inhibition of prostaglandins. However, this hypothesis was soon discounted on grounds that, if this had been the case, then one should have observed a decrease of asthma incidence when aspirin was first introduced (Shaheen et al. 2000). Thus the suspicion finally fell upon acetaminophen itself (Newson et al. 2000) and subsequent investigations explicitly aimed to examine the hypothesis of causal connection between acetaminophen and asthma. A series of observational and quasi-experimental studies have investigated the hypothesis that acetaminophen is causally associated with an increase of asthma incidence/severity (Newson et al. 2000; Lesko et al. 2002; Barr et al. 2004; McKeever et al. 2005; Karimi et al. 2006; Beasley et al. 2008; 2011b; Shaheen et al. 2008; Amberbir et al. 2011 see tables in appendix).

One among the most telling studies in this respect is a prospective survey realized over a sample of 73,000 female nurses (Barr et al. 2004), who were asked in 1990 and 1992 about their medication habits (acetaminophen included) and known diagnoses. In 1996, 346 among those of them who had not any record of asthma at the beginning of the study, were diagnosed with asthma. Women who took acetaminophen >14 days/month were 1.63 times as likely to be diagnosed with asthma as those who did not assume acetaminophen (95 % CI 1.11–2.39). The prospective design refuted the hypothesis of reverse causation, i.e. that asthma might induce a higher level of acetaminophen consumption, due to its clinical implications, such as higher than average frequency of fever and headache. Furthermore, the increase was dose-dependent and aspirin as well as nonsteroidal anti-inflammatory drugs showed little relationship to asthma.

The closest study to a RCT was a double blind randomized clinical study performed within the Boston University Fever Study. Among the children enrolled for the study (84,192, <12 y) a subgroup of subjects diagnosed with asthma (n = 1879) was evenly assigned to three distinct treatments consisting of low-dose ibuprofen, high-dose ibuprofen or acetaminophen (12 mg/kg per dose). A significant dose-dependent association was once again found between acetaminophen exposure and asthma exacerbation: those treated with acetaminophen for respiratory infection subsequently had a higher need of outpatient asthma visit (2.3 times higher; 95 % CI 1.26–4.16; Lesko et al. 2002). No dose-dependence for ibuprofen was found instead. Although short-term effects in already asthmatic subjects cannot be straightforwardly extrapolated to effects on asthma inception (Henderson and Shaheen 2013), these results provide significant supporting evidence for the existence of some mechanisms linking acetaminophen and asthma.

In fact, possible biological pathways mediating the causative action of acetaminophen for asthma had been identified well before a suspicion of causal connection between acetaminophen and asthma emerged in the epidemiological literature. Eneli et al. (2005) summarize these findings and present five possible (non-exclusive) causal pathways accounting for the role of acetaminophen in asthma exacerbation. Three pathways depend on the influence of acetaminophen on glutathione depletion. Glutathione molecules mitigate oxidative stress. Thus, by depleting glutathione, acetaminophen contributes to hyperresponsiveness to environmental antigens, thus promoting atopic disease; furthermore, it has been hypothesized that accumulation of the end product of acetaminophen metabolization (*N*-acetyl-p-benzoquinonemine) is conjugated by glutathione into a harmless substance: thus with insufficient presence of glutathione, *N*-acetyl-p-benzoquinonemine (NAPQI) accumulates in the liver and arylates cellular macromolecules, producing cell death; in this case, acetaminophen has a toxic effect by negatively impacting on the cellular "scavenger" which should clean up its toxic metabolite. Another possible glutathione dependent pathway of asthma exacerbation regards the switching, caused by reduced glutathione levels, from a T-helper-1 to a T-helper-2 (allergic) response to antigens.

One of the two pathways unrelated to glutathione involves the lack of suppression of cyclooxygenase, which is also an inflammatory pathway. Cyclooxygenase promotes the production of Prostaglandin E_2, which tilts the immunologic process towards a T2 response: thus, by inhibiting the suppression of cyclooxigenase, acetaminophen induces an allergic response and thereby increases the chance of asthma exacerbation or insurgence.

The fifth possible pathway is an immunologic response to acetaminophen itself: this possibility has been investigated through skin-prick tests which measured the level of acetaminophen-specific serum IgE after oral challenge of diverse quantities of acetaminophen (Galindo et al. 1998; de Paramo et al. 2000) and resulted to be positive. Furthermore, other studies have detected elevated levels of histamine in subjects treated with acetaminophen (histamine mediates the cascade of inflammation triggered by IgE). However, knowledge of the pathogenesis of other analgesia-induced asthma undermines the plausibility of this hypothesis, because this sort of effects is thought to be produced through the cyclooxigenase rather than through the IgE-mediated pathway.

More recently an additional immunologic pathway has been hypothesised, namely the production of neurogenic airway inflammation caused by the transient receptor potential ankyrin-1 (TRPA1): this is stimulated by detectable concentrations of NAPQI in the lung produced by acetaminophen (Nassini et al. 2010). Also, another possible acetaminophen-asthma mediating mechanism is linked to its antipyretic effect which possibly reduces the cytokine storm of febrile responses, thus reducing the level of Interferon-γ and Interleukin-2 and thereby predisposing the organism to an allergic (Th2) rather than to a non allergic response (Th1). However, whereas the antipyretic action of acetaminophen is well-established, its influence on cytokine production still needs to be proved and may depend on the cause of the fever (Farquhar et al. 2010). Finally, it has been shown that acetaminophen weakens the immune response to rhinovirus infection and prolongs it in volunteers (Graham et al. 1990).

Possible pathways are illustrated in Fig. 1.

While all pathways are only indirectly relevant to asthma pathogenesis, their plausibility is strongly supported by experimental data at different levels (in vitro, in vivo, and clinical studies). For some, this evidence provides some mechanistic rationale, and strengthens the support to the causal hypothesis provided by the evidence obtained at the population level, at the point that no additional randomized studies are needed in order to consider acetaminophen as a causative factor for asthma exacerbation or insurgence. Others instead hold a conservative view and are concerned by confounding. Indeed the *detection process* reflects the hierarchy reversal proposed by Vandenbroucke (2008): observational studies (and all the more, case reports and basic science) come earlier,[7] then comparative studies further investigate the causal relationship, finally prospective studies are meant to provide the guarantee that the causal association goes in the supposed direction. However *justification* of the causal hypothesis is far from reaching a consensus on this basis. In fact, the accruing evidence in favour of the acetaminophen-asthma connection, is generating two opposing stances in the scientific community.

7.2 The Acetaminophen Enigma in Asthma

The evidence gathered so far in support of the hypothesis of causal association between acetaminophen and asthma has generated two opposite stances. On one side, a series of authors show some reluctance in accepting such evidence as a sufficient basis for practice change and for establishing a causal relationship between acetaminophen and asthma, on grounds that it does not result from randomized clinical trials (Eneli et al. 2005; Allmers et al. 2009; Johnson and Ownby 2011; Karimi et al. 2006; Wickens et al. 2011; Chang et al. 2011). Particularly, these authors express the concern that the acetaminophen-asthma relationship may be explained by reverse causation, confounding by

[7] A case study suggested the association of acetaminophen and asthma as early as in 1967 (Chafee and Settipane 1967), but this has not triggered further analysis until the asthma epidemic of the 70's 90's.

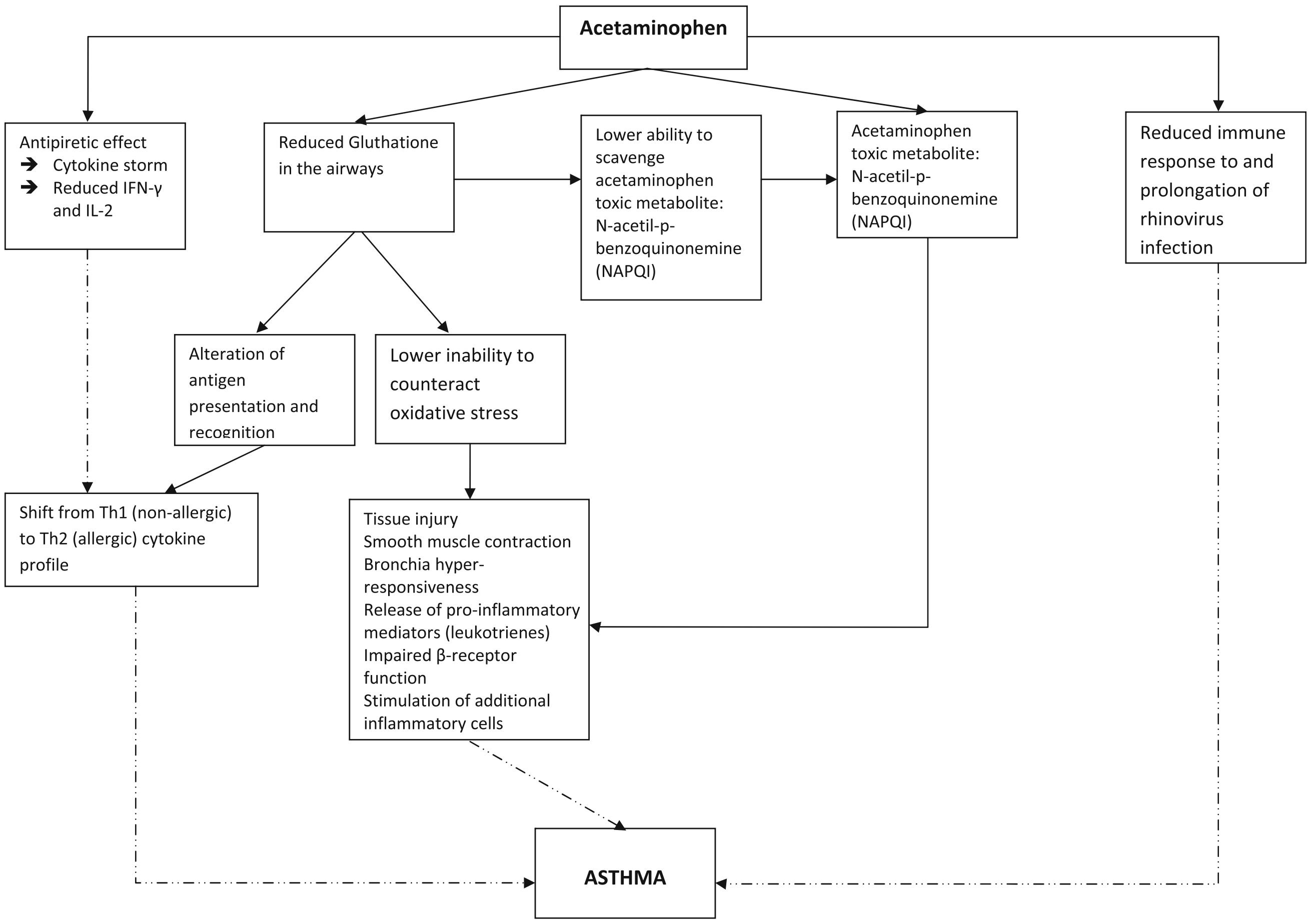
Acetaminophen
Antipiretic effect
➔ Cytokine storm
➔ Reduced IFN-γ and IL-2
Reduced Gluthatione in the airways
Lower ability to scavenge acetaminophen toxic metabolite: N-acetil-p-benzoquinonemine (NAPQI)
Acetaminophen toxic metabolite: N-acetil-p-benzoquinonemine (NAPQI)
Reduced immune response to and prolongation of rhinovirus infection
Alteration of antigen presentation and recognition
Lower inability to counteract oxidative stress
Shift from Th1 (non-allergic) to Th2 (allergic) cytokine profile
Tissue injury
Smooth muscle contraction
Bronchia hyper-responsiveness
Release of pro-inflammatory mediators (leukotrienes)
Impaired β-receptor function
Stimulation of additional inflammatory cells
ASTHMA

Fig. 1 Possible relevant pathways conducing from acetaminophen to asthma. *Solid arrows* represent established links (through in vivo or in vitro studies). *Dashed arrows* represent relevant associations

indication or preference for acetaminophen rather than ibuprofen in children at risk for asthma. Other authors, although less sceptical about the causal relationship, nevertheless equally require or recommend the performance of adequately powered placebo-controlled trials to establish causation (Holgate 2011; Henderson and Shaheen 2013). On the other side, Martinez-Gimeno and García-Marcos 2013, emphasize that "apart from tobacco smoke exposure, no other genetic or environmental factors, including genes, allergens, infections and bacterial substances, has shown the stubborn and consistent association with wheezing disorders prevalence as acetaminophen has done" (Martinez-Gimeno and García-Marcos 2013:120) and they recommend against a too liberal use of acetaminophen in children, while waiting for regulatory agencies to do their part and reconsider the safety profile of acetaminophen (Martinez-Gimeno and García-Marcos 2013:121). Furthermore they are against the performance of double blind RCTs with placebo, since "contrary to common claims, 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). Thus they recommend special kinds of RCTs, where the intervention is avoidance of acetaminophen (letting subjects being administered other analgesics) and control is free consumption of any analgesics, acetaminophen included.

According to other authors Beasley et al. (2011a), "*When the study findings are considered together with other available data*, there is substantive evidence that acetaminophen use in childhood may be an important risk factor for the development and/or maintenance of asthma, and that its widespread increasing use over the last 30 years may have contributed to the rising prevalence of asthma in different countries worldwide" (p. 1570, my emphasis). An even stronger commitment to the hypothesis of causal association is expressed by McBride (2011), who, considering all the evidence available, as well as his personal experience as a paediatrician pulmonologist, claims that evidence of causal association between acetaminophen and asthma can by now be regarded as strong enough to warrant a change in prescription practice. McBride justifies his claim by appealing to the consistency of interdisciplinary evidence: (1) strength of the association displayed in comparative studies; (2) robustness of association across geography, culture and age; (3) dose–response relationship between acetaminophen exposure and asthma; (4) coincidence of time trends in acetaminophen use and asthma increase; (5) lack of other equally strong causal explanations; (6) relationship between asthma epidemic and per-capita sales of acetaminophen across countries; (7) plausible mechanism. Particularly, because of the overlapping time trends of the asthma epidemic and of acetaminophen consumption, a causal link between the two seems to explain the asthma epidemic more than other environmental factors can do. McBride explicitly warns against the use of acetaminophen in children with asthma or at risk for asthma and claims that if further evidence is required, then this is for documenting product safety rather than the contrary, because its harmful potential is sufficiently demonstrated by the evidence collected so far. By shifting the burden of proof, McBride assumes that, given the available evidence, the hypothesis of causal connection between acetaminophen and asthma is stronger than that of its absence; or, at least, that given the expected harm and benefit, the probability of causal connection between acetaminophen and asthma is high enough as to shift the balance against its use.

However, whereas detractors of the causal association need nothing more than appealing to the received view that observational data cannot prove causation, supporters of the causal link between acetaminophen and asthma feel unease about the lack of RCTs confirming their views and mix up categories informing evidence hierarchies with other criteria of causal judgment such as Bradford Hill criteria or convergent evidence of different kinds, which are unrelated, and possibly in outright contrast with the very idea of ranking evidence. For instance, Martinez-Gimeno, García-Marcos' analysis (2013), explicitly draws on typical EBM categories, but then unwittingly blends them with reasoning about biological mechanisms, and with Bradford Hill criteria for causal assessment. Beasley et al. (2008) and McBride (2011) base their argumentation on convergent evidence of different kinds, but the latter feels compelled to call for a reversal of the burden of proof in order to substantiate his claim.

The dissent concerning the best course of action among scholars is ultimately caused by differing epistemological views which are left implicit. Those recommending the performance of placebo-controlled RCTs are in line with the rationales underlying evidence hierarchies. Thus they insist on the elimination of any suspicion of confounding, especially confounding by indication (Henderson and Shaheen 2013; Chang et al. 2011) before any causal claim can be established on firm grounds. On the other side, supporters of the causal link, especially McBride (2011) and Beasley et al. (2011a, b), point to the joint support of different and independent sources of evidence as a valid basis for dropping any need for RCTs. Who's right?

We might try to answer to this question by drawing on the three dimensions in the enterprise of causal assessment for harms mentioned in Sect. 6: (1) integration of prior knowledge and available evidence; (2) cumulative causal learning and probabilistic assessment of causality; (3) impartiality.

1. In the acetaminophen case, prior knowledge about the molecule itself would be rather against the hypothesis of harmfulness, in that it has been generally considered an harmless analgesics, and this might also explain the reluctance to accept this causal hypothesis (Martinez-Gimeno and García-Marcos 2013). However many *prima facie* harmless substances have been retired from the market after discovering surprising noxious effects. Furthermore, in the case of acetaminophen a relevant amount of biological data point to its potential inflammatory effects on the airways through multiple (possibly additive) pathways. Dismissal of the causal link because of possible confounding factors at the epidemiological level explicitly eludes this evidence. This is also valid for other supporting evidence such as the dose–response relationship found in many studies, and in general for the higher likelihood of the entire set of data on the hypothesis of causation rather than on its denial;
2. Whereas detractors of the causal hypothesis seem to feel uncommitted until contrary proven and advocate for the performance of RCTs before taking any action, its supporters feel challenged by the evidence already available and consider what should be thought and done on its basis. Contrary to what is expected, the former attitude is not neutral since its default is that that there is no causal association, until proved by RCTs, whereas the available evidence does no longer warrant the categorical rejection of this hypothesis;
3. Dismissal of the causal association between acetaminophen and asthma on grounds that the overwhelming epidemiological evidence may be produced by confounders represents a case where uncertainty about causal connection may be exploited by interested parties (Lowe et al. 2010 and Holgate 2011 have conflicting interests for instance). In the end, a too rigid attitude towards evidence quality may run against the reasons for which quality standards have been introduced.

8 Conclusion

Both philosophers as well as epidemiologists and health scientists have acknowledged the role of so called "lower level" evidence as a valid source of information contributory to assessing the risk profile of medications on theoretical (Aronson and Hauben 2006; Howick et al. 2009; Vandenbroucke 2006) or empirical grounds (Papanikolaou et al. 2006; Benson and Hartz 2000; Golder et al. 2011; Concato et al. 2000). However, they have difficulty in assigning a precise epistemic status to this kind of evidence and in providing a coherent justification for the sorts of exemptions they propose. In general, evidence quality coming from other sources than RCTs is considered high to the extent that possible known or unknown confounders can be safely excluded (Glasziou et al. 2007; Howick 2011; Howick et al. 2009). Vandenbroucke (2008) adds to this scheme the distinction between the context of evaluation and the context of discovery as a possible explanation for the asymmetry between evidence standards for benefit versus risk assessment. On this basis he also proposes to reverse evidence hierarchies when risks rather than benefits of medical treatment are under scrutiny. (Psaty and Vandenbroucke 2008). This is only part of the issue however.

My analysis has focused on the distinction between "clinchers" and "vouchers" (Cartwright 2007) intended as two opposite stances towards scientific inference and evidence evaluation. Methods such as hypothesis testing are clinchers in that they are based on deductive rules of inference; instead inductive and abductive methods of hypothesis assessment are vouchers in that they cannot guarantee the conclusion which they favour. Whereas clinchers work with truth conditions and put severe constraints on the kind of evidence which can be admitted, or at all considered meaningful in assessing causal associations; inductive and abductive methods put different sorts of evidence together and infer the implications which derive from their joint support to the hypotheses under investigation. Roughly speaking, the former methods aim to hypothesis rejection or acceptance with no degree in-between, while the latter aim to a judgment on the plausibleness of the hypothesis given the data, which can be possibly quantified in probabilistic terms.

Evidence hierarchies are based on clinchers and ranking aims to internal validity maximisation thus promoting a "take the best" approach. Instead vouchers work with the principle of total evidence. Thus the reasons for accepting "lower level evidence" differ in the two settings. But current practices have difficulty in assigning a precise epistemic status to this kind of evidence because they more or less implicitly stick to the rationales underpinning evidence hierarchies, has illustrated by the acetaminophen case.

The tension between detractors and supporters of the necessity to perform placebo-controlled RCTs before establishing a causal link between acetaminophen and asthma originates exactly from the antagonism between two school of thoughts, clinchers enthusiasts versus vouchers adherents, which is left implicit. I unearthed the different epistemic paradigms underlying these different methodological stances in order to (1) allow a transparent discussion of the reasons why "lower level evidence" may/should be accepted in each specific context; (2) provide a theoretical underpinning to the increasing consensus that evidence for harms should be evaluated with different standards than those used for testing benefit claims.

Acknowledgments Thanks go to two anonymous referees and to: Nancy Cartwright, Jeremy Howick, Phyllis Illari, Bert Leuridan, David Papineau, Julian Reiss, Federica Russo, David Teira, and Jan Vandenbroucke. All provided precious comments on earlier drafts of the paper. I am also indebted to the audiences to which the paper has been presented at different stages of the research project: Evidence and Causality Conference (Canterbury); Conference of the Italian Society for Analytical Philosophy (Alghero); IFOM-IEO Campus (Milan); Concepts of Health Seminar at King's College London. Funding for the paper has been provided by an intramural project conducted with Prof. Fiorenzo Mignini at the University of Camerino: "Epistemic Asymmetries in Benefit vs. Risk Assessment of Pharmaceuticals".

References

Agency for Healthcare Research and Quality (2007) Methods reference guide for effectiveness and comparative effectiveness reviews, Version 1.0 [Draft posted October 2007). Available at: http://www.effectivehealthcare.ahrq.gov/repFiles/2007_10DraftMethodsGuide.pdf

Allmers H, Skudlik C, John SM (2009) Acetaminophen use: a risk for asthma? Curr Allergy Asthma Rep 9(2):164–167

Amberbir A, Medhin G, Alem A, Britton J, Davey G, Venn A (2011) 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 183(2):165–170

Anderson HR (1989) Increase in hospital admissions for childhood asthma: trends in referral, severity and readmissions from 1970 to 1985 in a health region of the United Kingdom. Thorax 44:614–619

Aronson JK, Hauben M (2006) Anecdotes that provide definitive evidence. BMJ 333(16):1267–1269

Ayer AJ (1956) The problem of knowledge. MacMillan, London

Ayer AJ (1957) The conception of probability as a logical relation. In: Korner S (ed) Observation and interpretation in the philosophy of physics. Dover Publications, New York

Barr RG, Wentowski CC, Curhan GC et al (2004) Prospective study of acetaminophen use and newly diagnosed asthma among women. Am J Respir Crit Care Med 169(7):836–841

Basu D (1980) Randomization analysis of experimental data: the fisher randomization test. J Am Stat Assoc 75(371):593–595

Beasley RW, Clayton TO, Crane J, ISAAC Phase Three Study Group et al (2008) Association between acetaminophen 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 372(9643):1039–1048

Beasley RW, Clayton TO, Crane J, ISAAC Phase Three Study Group et al (2011a) 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 183(2):171–178

Beasley RW, Crane J, Lai C, Stewart A, Clayton T (2011b) Acetaminophen and asthma: spurious association? Am J Respir Critical Care Med; Jun 1;183(11): author reply 1570–1

Benson K, Hartz AJ (2000) A comparison of observational studies and randomised, controlled trials. N Engl J Med 342(25):1878–1886

Broadbent A (2011) Inferring causation in epidemiology: mechanisms, black-boxes, and constraints. In: Illari PM, Russo F, Williamson J (eds) Causality in the Sciences. Oxford University Press, Oxford

Burr ML, Butland BK, King S, Vaughan Williams E (1989) Changes in asthma prevalence: two surveys 15 years apart. Arch Dis Child 64:1452–1456

Carnap R (1947) On the application of inductive logic. Philos Phenomenol Res 8(1):133–148

Cartwright N (2007) Are RCTs the gold standard? Biosocieties 2:11–20

Cartwright N (2010) Presidential address. Am Philos Assoc Pac Div, Vancouver, April 10

Chafee FH, Settipane GA (1967) Asthma caused by FD&C approved dyes. J Allergy 40:347–351

Chalmers I (2005) Statistical theory was not the reason that randomization was used in the British Medical Research Council's clinical trial of streptomycin for pulmonary tuberculosis. In: Jorland G, Opinel A, Weisz G (eds) Body counts: medical quantification in historical and sociological perspectives. McGill-Queens University Press, Montreal, pp 309–334

Chang KC, Leung CC, Tam CM, Kong FY (2011) Acetaminophen and asthma: spurious association? Am J Respir Crit Care Med 183(11):1570

Clarke B, Gillies D, Illari P, Russo F, Williamson J (2012) The evidence that evidence-based medicine omits. Prev Med. doi:10.1016/j.ypmed.2012.10.020

Clarke B, Gillies D, Illari P, Russo F, Williamson J (forthcoming) Mechanisms and the evidence hierarchy. TOPOI, special issue "Evidence and Causality in the Sciences"

Concato J, Shah MPH, Horwitz RI (2000) Randomized, controlled trials, observational studies and the hierarchy of research designs. N Engl J Med 342(25):1887–1892

de Paramo BJ, Gancedo SQ, Cuevas M et al (2000) Acetaminophen (acetaminophen) hypersensitivity. Ann Allergy Asthma Immunol 85:508–511

Eder W, Ege JM, von Mutius E (2006) The asthma epidemic. NEJM 355(21):2226–2235

Eneli I, Sadri K, Camargo C, Barr RG (2005) Acetaminophen and the risk of asthma: the epidemiologic and pathophysiologic evidence. CEST 127(2):604–612

Evidence Based Medicine Working Group (1992) Evidence-based medicine: a new approach to the teaching the practice of medicine. JAMA 268(17):2420–2425

Farquhar H, Stewart A, Mitchell E, Crane J, Eyers S, Weatherall M, Beasley R (2010) The role of paracetamol in the pathogenesis of asthma. Clin Exp Allergy 40(1):32–41

Freedman B (1987) Equipoise and the ethics of clinical research. N Engl J Med 317:141–145

Galindo PA, Borja J, Mur P et al (1998) Anaphylaxis to acetaminophen. Allergol Immunopathol 26:199–200

Gifford F (1995) Community equipoise and the ethics of randomized clinical trials. Bioethics 9:127–148

Gifford F (2007a) So-called 'equipose' and the argument form design. J Med Philos 32:135–150

Gifford F (2007b) Taking equipoise seriously: The failure of clinical or community equipoise to resolve ethical dilemmas in randomized clinical trials. In: Kincaid H, McKitrick J (eds) Establishing medical reality essays in the metaphysiscs and epistemology of biomedical science. Springer, New York

Glasziou P, Chalmers I, Rawlins M, McCullock P (2007) When are randomized trials necessary? Picking signal from noise. Br Med J 334:349–351

Golder S, Loke YK, Bland M (2011) Meta-analyses of adverse effects data derived from randomized controlled trials as compared to observational studies: methodological overview. PLoS Med 8(5):1–13

Graham NM, Burrell CJ, Douglas RM, Debelle P, Davies L (1990) Adverse effects of aspirin, acetaminophen, and ibuprofen on immune finction, viral shedding, and clinical status in rhinovirus-infected volunteers. J Infect Dis 162(6):1277–1282

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 (2011) GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 64(4):383–394

Guyatt GH, Oxman AD, Vist GE et al (2008) GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. Br Med J 336:924–926

Henderson AJ, Shaheen SO (2013) Acetaminophen and asthma. Paediatr Respir Rev 14(1):9–16

Herxheimer A (2012) Pharmacovigilance on the turn? Adverse reactions methods in 2012. Br J Gen Pract 62(601):401–402

Holgate ST (2011) The acetaminophen enigma in asthma. Am J Respir Crit Care Med 183:147–148

Howick J (2011) The Philosophy of Evidence-Based Medicine. Wiley-Blackwell, NY

Howick J, Glasziou P, Aronson JK (2009) The evolution of evidence hierarchies: what can Hill's 'guidelines for causation' contribute? J R Soc Med 102:186–194

Howick J, Chalmers I, Glasziou P, Greenhalgh T, Heneghan C, Liberati A, Moschetti I, Phillips B, and Thornton H (2011) The 2011 Oxford CEBM evidence levels of evidence (introductory document). Oxford Centre for Evidence-Based Medicine. http://www.cebm.net/index.aspx?o=5653

Howick J, Glasziou P, Aronson JK (2013) Can understanding mechanisms solve the problem of extrapolating from study to target populations (the problem of 'external validity')? J R Soc Med 106:81–86

Howson C, Urbach P (2006) Scientific reasoning: the Bayesian approach. Open Court, Illinois

Jick H (1977) The discovery of drug-induced illness. N Engl J Med 296:481–485

Jick H, Vessey M (1978) Case-control studies in evaluation of drug-induced illness. Am J Epidemiol 107:1–7

Joffe M (2011) Causality and evidence discovery in epidemiology. In: Dieks D, Gonzales WJ, Hartmann S, Uebel T, Wever M (eds) Explanation, prediction, and confirmation. Springer, Berlin, pp 153–166

Johnson CC, Ownby DR (2011) Have the efforts to prevent aspirin-related Reye's syndrome fuelled an increase in asthma? Clin Exp Allergy 42(3):296–298

Karimi M, Mirzaei M, Ahmadieh MH (2006) Acetaminophen use and the symptoms of asthma, allergic rhinitis and eczema in children. Iran J Allergy Asthma Immunol 5(2):63–67

Keynes JM (1921) A treatise on probability. MacMillan, London

Kyburg HE, Teng CM (2001) Uncertain inference. Cambridge University Press, Cambridge

La Caze A (2009) Evidence based medicine must be. J Med Philos 34(5):509–527

La Caze A (2013) Why randomized interventional studies. J Med Philos 38(4):352–368

Lesko SM, Louik C, Vezina RM, Mitchell AA (2002) Asthma morbidity after the short-term use of ibuprofen in children. Pediatrics 109(2):e20. Available at http://pediatrics.aappublications.org/content/109/2/e20.full.pdf+html

Lindley DV (1982) The role of randomization in inference. PSA: Proceedings of the Biennial meeting of the philosophy of science association, vol II: symposia and invited papers, pp 431–446

Lowe A, Carlin JB, Bennett CM, Hosking CS, Allen KJ, Robertson CF, Axelrad C, Abramson MJ, Hill DJ, Dharmage SC (2010) Paracetamol use in early life and asthma: prospective birth cohort study. BMJ 341:c4616

Mannino DM, Homa DM, Pertowski CA, Ashizawa A, Nixon LL, Johnson CA, Ball LB, Jack E, Kang DS (1998) Surveillance for Asthma—United States, 1960–1995. MMWR CDC Surveill Summ 47(SS-1):1–28

Mannino DM, Homa DM, Akinbami LJ, Moorman JE, Gwynn C, Redd SC (2002) Surveillance for Asthma—United States, 1980–1999. MMWR CDC Surveill Summ 51(SS01):1–13

Martinez-Gimeno A, García-Marcos L (2013) The association between acetaminophen and asthma: should its pediatric use be banned? Expert Rev Respir Med. 7(2):113–122

McBride JT (2011) The association of acetaminophen and asthma prevalence and severity. Pediatrics 128(6):1181–1185. Available at: http://pediatrics.aappublications.org/content/128/6/1181.full.pdf+html?sid=29b845ce-0023-48dd-9835-3c87a6b45b69

McKeever TM, Lewis SA, Smit HA, Burney P, Britton JR, Cassano PA (2005) The association of acetaminophen, aspirin, and ibuprofen with respiratory disease and lung function. Am J Respir Crit Care Med 171(9):966–1071

Meheus J (2011) A formal logic for the abduction of singular hypotheses. In: Dieckes D, Gonzales WJ, Hartmann S, Uebel T, Weber M (eds) Explanation, Prediction, and Confirmation. Springer

Miettinen OS (1983) The need for randomization in the study of intended effects. Stat Med 2:267–271

Nassini R, MAterazzi S, Andre E, Sartiani L, Aldini G, Trevisani M et al (2010) Acetaminophen, via its reactive metabolite N-acetyl-p-benzo-quinonemine and transient receptor potential ankyrin-1 stimulation, causes neurigenic inflammation in the airways and other tissues in rodents. FASEB J 24:4904–4916

Newson RB, Shaheen SO, Chinn S, Burney PG (2000) Acetaminophen sales and atopic diseases in children and adults: an ecological analysis. Eur Respir J 16(5):817–823

Ninan TK, Russell G (1992) Respiratory symptoms and atopy in Aberdeen schoolchildren: evidence from two surveys 25 years apart. Br Med J 304:873–875

Osimani B (2007) Probabilistic information and decision-making in the health context: Package leaflets as basis for informed consent. PhD Thesis, University of Lugano

Osimani B (2013a) The precautionary principle in the pharmaceutical domain: a philosophical enquiry into probabilistic reasoning and risk aversion. Health, Risk & Society (special issue: Health Care through the Lens of Risk) 15(2):123–143

Osimani B (2013b) Until RCT-proven? On the asymmetry of evidence requirements for risk assessment. J Eval Clin Pract 19:454–462

Osimani B, Russo F, Williamson J (2011) Scientific evidence and the law: an objective Bayesian formalisation of the precautionary principle in pharmaceutical regulation. J Philos Sci Law 11

Papanikolaou PN, Christidi GD, Ioannidis JPA (2006) Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. Can Med Assoc J 174(5):635–641

Papineau D (1994) The virtues of randomization. Br J Philos Sci 45(2):437–450

Platts-Mills TA, Erwin E, Heymann P, Woodfolk J (2005) Is the hygiene hypothesis still a viable explanation for the increased prevalence of asthma? Allergy 60(suppl 79):25–31

Psaty B, Vandenbroucke JP (2008) Opportunities for enhancing the FDA guidance on pharmacovigilance. JAMA 300(8):952–953

Russo F, Williamson J (2007) Interpreting causality in the health sciences. Int Stud Philos Sci 21(2):157–170

Russo F, Williamson J (2011) Epistemic causality and evidence-based medicine. Hist Philos Life Sci 33(4):563–582

Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS (1996) Evidence based medicine: what it is and what it isn't. Br Med J 312:71–72

Seaton A, Godden DJ, Brown K (1994) Increase in Asthma: a more toxic environment or a more susceptible population? Thorax 49:171–174

Shaheen SO, Sterne JAC, Songhurst CE, Burney PGJ (2000) Frequent acetaminophen use and asthma in adults. Thorax 55:266–270

Shaheen S, Potts J, Gnatiuc L et al (2008) Selenium and Asthma Research Integration Project; GA2LEN. The relation between acetaminophen use and asthma: a GA2LEN European case-control study. Eur Respir J 32(5):1231–1236

Stegenga J (2011) Is meta-analysis the platinum standard of evidence? Stud Hist Philos Biol Biomed Sci 42:497–507

Straus SE, McAlister FA (2000) Evidence-based medicine: a commentary on common criticisms. Can Med Assoc J 163(7): 837–841

Stricker B, Psaty B (2004) Detection, verification, and quantification of adverse drug reactions. Br Med J 329:44–47

Subbarao P, MAndhane PJ, Sears MR (2009) Asthma: epidemiology, etiology, risk factors. CMAJ 181(9):181–190

Suppes P (1982) Arguments for randomizing. PSA 2:464–475

Teira D (2011) Frequentist versus Bayesian Clinical Trials. In: Gifford F (eds) Handbook of the philosophy of science, vol 16: Philosophy of Medicine pp 255–298

Teira D, Reiss J (2013) Causality, impartiality and evidence-based policy. In: Chao H-K, Chen S-T, Millstein R (eds) Towards the Methodological Turn in the Philosophy of Science: Mechanism and Causality in Biology and Economics, Springer, New York, pp 207–224

Urbach P (1994) Reply to David Papineau. Br J Philos Sci 45(2):712–715

Vandenbroucke JP (2004) When are observational studies as credible as randomized trials? Lancet 363:1728–1731

Vandenbroucke JP (2006) What is the best evidence for determining harms of medical treatment? Commentary. Can Med Assoc J 174(5):645–646

Vandenbroucke JP (2007) Letter to the editor. JAMA 297(19): 2077–2078

Vandenbroucke JP (2008) Observational research, randomised trials, and two views of medical science. Plos Med 5(3):339–343 (quote in the text are from the longer version: 1–28)

Vandenbroucke JP, Psaty BM (2008) Benefits and risks of drug treatments. How to combine the best evidence on benefits with the best data about adverse effects. JAMA 300(20):2417–2419

Varner AE, Busse WW, Lemanske RF Jr (1998) Hypothesis: decreased use of pediatric aspirin has contributed to the increasing prevalence of childhood asthma. Ann Allergy Asthma Immunol 81(4):347–351

Wickens K, Beasley R, Town I, Epton M, Pattemore P, Ingham T, Crane J; New Zealand Asthma and Allergy Cohort Study Group. (2011) The effects of early and late acetaminophen exposure on asthma and atopy: a birth cohort. Clin Exp Allergy. 2011 Mar; 41(3):399–406. doi:10.1111/j.1365-2222.2010.03610.x. Epub 2010 Sep 29

Worral J (2007a) Evidence in medicine and evidence-based medicine. Philos Compass 2(6):981–1022

Worral J (2007b) Why there's no cause to randomize? Br J Philos Sci 58:451–488

Worral J (2010) Do we need some large, simple randomized trials in medicine? In: Suárez, Mauricio and Dorato, Mauro and Rédei, Miklós, (eds.) EPSA philosophical issues in the sciences: launch of the European Philosophy of Science Association. Springer, London