Preventive Medicine Reports 1 (2014) 9–13

Contents lists available at ScienceDirect

Preventive Medicine Reports

journal homepage: http://ees.elsevier.com/pmedr

Commentary

Safety vs. efficacy assessment of pharmaceuticals: Epistemological rationales and methods

Barbara Osimani

University of Camerino, School of Pharmacology, P.zza dei Costanti, 62032 Camerino, Italy

ARTICLE INFO

Available online 26 August 2014

Keywords:
Safety assessment
Adverse drug reactions
Evidence-based medicine
Evidence hierarchies
Causal assessment
Hypothesis testing
Inductive methods
Bayesian epistemology

ABSTRACT

In their comparative analysis of Randomised Clinical Trials and observational studies, Papanikoloau et al. (2006) assert that "it may be unfair to invoke bias and confounding to discredit observational studies as a source of evidence on harms". There are two kinds of answers to the question why this is so. One is based on metaphysical assumptions, such as the problem of causal sufficiency, modularity and other statistical assumptions. The other is epistemological and relates to foundational issues and how they determine the constraints we put on evidence. I will address here the latter dimension and present recent proposals to amend evidence hierarchies for the purpose of safety assessment of pharmaceuticals; I then relate these suggestions to a case study: the recent debate on the causal association between paracetamol and asthma. The upshot of this analysis is that different epistemologies impose different constraints on the methods we adopt to collect and evaluate evidence; thus they grant "lower level" evidence on distinct grounds and at different conditions. Appreciating this state of affairs illuminates the debate on the epistemic asymmetry concerning benefits and harms and sets the basis for a foundational, as opposed to heuristic, justification of safety assessment based on heterogeneous evidence.

Introduction

Evidence standards are supposed to provide "quick and dirty rules" for assessing the quality of evidence, as a function of the greatest possible reduction of confounding and bias. Hence, 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 pathophysiologic 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 philosophical overview).

At present, no distinction is explicitly made concerning the role of such standards for assessing safety instead of efficacy. However, recent contributions by philosophers 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. Some of them are based on empirical surveys attesting that both randomized clinical studies and observational ones deliver the same incidence estimates for a series of risks associated with medical interventions, both pharmaceutical and surgical (Benson and Hartz, 2000; Golder et al., 2011; Papanikolaou et al., 2006). Others draw on various methodological considerations (Aronson and Hauben, 2006; Howick et al., 2009; Vandenbroucke, 2008). These suggestions have noteworthy implications when considering current emphasis on evidence hierarchies, since they imply an asymmetry in the way evidence of benefits and risks of health technologies should be evaluated. However such suggestions fail to be grounded on a sound epistemic basis and seem rather ad hoc, although intuitively correct. Thus my aim here is to present their epistemological underpinnings by relating them to the common partition into (statistical) hypothetico-deductive, abductive, and inductive(-Bayesian) approaches to scientific inference. Furthermore, I will point out a series of pragmatic constraints for which inductive rather than deductive approaches to scientific inference should be considered as better suited to the purpose of risk assessment.

The rationales underpinning evidence hierarchies and alternative approaches: necessary vs. uncertain inference

Scientific inference may be categorized in two main typologies: on one side, inferences whose conclusion follows necessarily from the set of premises and "rules" involved (e.g. laws and initial conditions). On the other side, cases falling under the heading of "uncertain inference", i.e. situations in which the conclusion is highly probable or plausible, but fallacious from a strictly logical point of view. The former generally fall into the category of deductive inference, whereas the latter are generally labeled with the umbrella category of "induction". An additional form of uncertain inference falls under the heading of inverse induction

E-mail addresses: barbaraosimani@gmail.com, barbara.osimani@unicam.it.

http://dx.doi.org/10.1016/j.pmedr.2014.08.002

(or abduction, as proposed by C.S. Peirce).[1] Both induction and abduction are rationally compelling but logically unwarranted methods of hypothesis confirmation. The distinction between them mainly consists in the former relying on probabilistic evidence, whereas the latter focuses on explanatory considerations. Current theories of scientific justification have coined an alternative term for abduction; this is "Inference to the Best Explanation" (also known as IBE, see Lipton, 2004); although not completely overlapping, the two concepts heavily rely on theoretical virtues such as simplicity/parsimoniousness (Ockham's razor) and informativeness/explanatory power as (imperfect indicators) of reliability: explanatory power of a theory is a mark of its truth. Deductive and inductive/abductive rationales underpin diverse methods of evidence evaluation for causal assessment as illustrated below.

Hypothesis testing as a deductive approach to scientific inference

In classical hypothesis-testing, the result is expressed as the probability of observing the experimental result or more "extreme" results in the sample space (p-value), if the treatment makes no difference (so called null Hypothesis: H_0). The underlying epistemology is hypothetico-deductive (Popper, 1992): one assumes an entailment relationship between lack of efficacy and lack of difference between treated and untreated group ($H_0 \rightarrow E$). If the outcome shows a statistically significant difference ($\neg E$), then the hypothesis of lack of treatment can be rejected ($\neg H_0$), following classical modus tollens:

$$\frac{H_0 \rightarrow E \quad ; \quad \neg E}{\neg H_0.}$$

In order to be able to draw a causal inference from the observed result, one must be confident that the difference between the two comparison groups is due to the contribution of the investigated factor, and only to it, otherwise ¬E might be due (also) to some alternative cause. Blinding, intervention and randomization are essential instruments in warranting this causal claim (see also Papineau, 1993; Worral, 2007; Osimani, 2013a,b,c) and evidence hierarchies are based on such warrants of internal validity. The EBM paradigm has been developed as a way to meet the desideratum that efficacy should be evaluated on the basis of the "best evidence" available, where "best" refers to quality criteria mainly informed by the requirement of internal validity.

The focus on internal validity is evident also in allowed deviations from evidence hierarchies in specific cases, i.e. where "lower level evidence" such as case reports and observational data are considered sufficient evidence for causal claims *to the extent that other conditions warrant for lack of bias and confounding*as alternative to randomization, blinding and intervention. 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 if studies are based on just observational evidence: these are represented as phenomena of sudden and drastic changes 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. Vandenbroucke (2008) considers that unintended effects, qua unintended, are not known in advance, and thus also not known by the drug prescriber, who cannot calculate on them and thereby possibly bias treatment allocation. It follows that observational studies concerning adverse reactions do not suffer from confounding in the same way as observational studies for intended effects do.

[1] Peirce introduced the term "abduction" first in "Deduction, Induction and Hypothesis" (1934, Collected Papers 2.623), then in the Cambridge Conferences (1898) and in the 1903 Harvard Lectures (with different semantic nuances: see also Thagard, 1988, Section 4.2.1).

Uncertain inference: probabilistic and explanatory approaches

Non-deductive methods abandon the goal of outright hypothesis acceptance or rejection and track uncertainty while updating the degree of confidence in a given hypothesis upon new evidence by also taking into account background knowledge. This allows them to be more flexible with regard to the kind of evidence which is allowed to inform hypothesis confirmation and the methods for amalgamating it.

Within this framework, the two somewhat contending paradigms are constituted by probabilistic approaches to hypothesis confirmation (e.g. Bayesian epistemology) and abductive reasoning (also fleshed out as "inference to the best explanation", Lipton, 2004).

Bayesian epistemologies (Howson and Urbach, 2006) insist on hypothesis confirmation rather than testing, and allow statistics to measure the degree of confirmation provided by evidence E to a given set of hypotheses $H = \{h_1, \ldots, h_n\}$, by relying both on the likelihood of the evidence in relation to each hypothesis $P(E / h_i)$, as well as on the probability measure associated to each hypothesis prior to collecting the evidence, $P(h_i)$, and by updating it through conditionalization (or other means, depending on the specific Bayesian approach adopted). This distinguishes them sharply from frequentist statistics where the p-value measures instead the probability of observing the evidence obtained in the experiment (or "more extreme results") if the hypothesis under investigation is false.

In the Bayesian paradigm the main requirement is that all available evidence is used (Carnap, 1947; Carnap, 1950): this is because all non-deductive logics are non-monotonic. Non-monotonicity is a phenomenon which characterizes defeasible reasoning, i.e. contexts where the addition of further data to the initial premises may invalidate some previous conclusion (Kyburg and Teng, 2001): formally, you may for instance have that the probability of hypothesis H is greater than its negation given evidence E: $P(H/E) > P(\neg H/E)$; but by adding another datum to the previous body of evidence, the opposite inequality may hold: $P(H/E,F) < P(\neg H/E,F)$. This may be illustrated by a diagnosis of celiac disease (H) with evidence of immune reactions to certain kinds of food (E), and then weakening of this hypothesis after a laboratory test (F).

In the IBE paradigm, hypotheses are justified by their explanatory power: the greater the amount of data the hypothesis is able to explain, the greater its plausibility. Thus, explanatory power is considered to be truth-conducive (Lipton, 2004). This paradigm is seldom explicitly adopted in causal assessment for health technologies; however it often underlies systematic reviews and qualitative reports, where heterogeneous evidence is combined in a narrative fashion.

The first important advocate of alternative approaches to statistical hypothesis testing was Sir Austin Bradford Hill with his most cited President's Address (Hill, 1965) inaugurating the Section of Occupational Medicine of the Royal Society of Medicine; that is, a discipline mostly concerned with exposure to hazards. After presenting his nine guidelines for detecting and assessing causal relationships he claims: "None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a sine qua non. What they can do, with greater or less strength, is to help us make up our minds in the fundamental question — *is there any other way of explaining the set of facts before us, is there any other equally, or more, likely than cause and effect*?" (emphasis added). Thus, Hill both refers to explanatory power and hypothesis likelihood as reliable grounds to justify causal judgments.

In recent times other authors have endorsed similar proposals. Aronson and Hauben (2006) put forward that "In some cases other types of evidence may be more useful than a randomized controlled trial. Combining randomized trials with observational studies and case series can sometimes yield information that is not available from randomized trials alone". This idea is also at the basis of the recent proposal by Howick et al. (2009) to integrate evidence hierarchies with Bradford Hill criteria for causal inference (see also Stegenga, 2011). Vandenbroucke (2008) proposes to invert hierarchies for