Chapter 15
Epistemic Gains and Epistemic Games: Reliability and Higher Order Evidence in Medicine and Pharmacology

Barbara Osimani

Abstract In this paper I analyse the dissent around evidence standards in medicine and pharmacology as a result of distinct ways to address epistemic losses in our game with nature and the scientific ecosystem: an "elitist" and a "pluralist" approach. The former is focused on reliability as minimisation of random and systematic error, and is grounded on a categorical approach to causal assessment, whereas the latter is more focused on the high context-sensitivity of causation in medicine and in the soft sciences in general, and favours probabilistic approaches to scientific inference, as better equipped for defeasibility of causal inference in such domains. I then present a system for probabilistic causal assessment from heterogenous evidence that makes justice of concerns from both positions, while also incorporating "higher order evidence" (evidence/information about the evidence itself) in hypothesis confirmation.

15.1 Introduction

Medical science distinguishes itself with respect to other sciences and technologies by the joint interaction of following phenomena:

1. Epistemic uncertainty with respect to the real state of nature and of the outcome of interventions is generally higher than for other natural sciences (Joffe 2011),

Funding by the European Research Council through grant 639276 (PhilPharm) is grateful acknowledged.

B. Osimani (✉)
Department of Biomedical Sciences and Public Health, Polytechnic University of the Marche, Ancona, Italy

Munich Center for Mathematical Philosophy, Munchen, LMU, Munich, Germany
e-mail: b.osimani@staff.univpm.it

A. LaCaze, B. Osimani (eds.), *Uncertainty in Pharmacology*, Boston Studies in the Philosophy and History of Science 338,
https://doi.org/10.1007/978-3-030-29179-2_15

indeed among the natural sciences, medicine is the closest one to the social sciences and the humanities;

2. Medical products are so called “credence products”, that is products for which the consumer (medical community, patients, and public health system in a more general sense) cannot evaluate the quality prior (and often not even after) consumption. Such state of affairs is therefore characterized by information asymmetry, which further interacts with the decision-makers’ risk attitude.
3. medical choices are high-stakes decisions both in terms of physical, psychological, existential, and financial costs (Osimani 2012),
4. information asymmetry and high sensitivity of medical decisions, can be strategically exploited by producers of medical knowledge having vested interests in the research outputs and dissemination (such strategic behaviour may of course also evolve in time) (Teira 2011; Teira and Reiss 2013; Holman 2015);
5. medical decisions are intertwined with pressing ethical dilemmas touching different anthropological dimensions (Papa 2014; Sgreccia 2007; Pessina 2009; Beauchamp 2011; Faden and Beauchamp 1986; Scheu 2003);

This picture is especially vivid in pharmacology; the complex network of interests (financial issues, concern for reputation etc.), as well as legal rights and duties which frame the scientific and social ecosystem in which pharmacology is embedded make it a unique blend of science and technology. Indeed, it is manifest that in scientific domains characterised by vested interests, the production and evaluation of evidence is embedded in a strategic game, where agents obviously try to maximise their payoff. This state of affairs strongly emphasises the role of reliability (in its various aspects) as a decisive dimension of evidence.

This paper focuses on the roots of methodological dissent in medicine as a dissent related to how reliability is conceptualised and warranted in contending schools of thoughts.

The paper is organised as follows: in Sect. 15.2, I present the debate around standards of evidence in medicine and the role played by evidence from randomized controlled trials vs. evidence of biological mechanisms or other kinds of sources. Section 15.3 presents the two poles on which such debate rests: the “elitist game” and the “pluralist game”. These two distinct approaches focus on different aspects of reliability as a result of weighting the costs of different kinds of errors differently. In particular, elitists strive to maximise internal validity by minimising random and systematic error, while the pluralists are rather concerned by the high context-sensitivity of causes in the biological realm (and in the soft sciences in general), hence they are more concerned about the stability of causal knowledge across populations and domains (failure of external validity and false predictions). In Sect. 15.4, I illustrate a meta-level perspective where the *structure* and *organisation* of the entire body of evidence brings its own contribution to hypothesis confirmation. In this perspective, epistemic dimensions of evidence itself, such as its reliability, consistency/coherence and (in)dependence also contribute to hypothesis confirmation in an interactive way, and these are connected with research organisation issues: how the scientific ecosystem is modelled and interacts with the broader social system. This perspective is concretely translated in a multilayer framework which models probabilistic causal inference through evidence synthesis elaborated by De Pretis et al. (2019). This framework, henceforth “E-Synthesis”, reconciles various concerns around causal inference in medicine and pharmacology, by allowing various kinds of evidence to jointly contribute to hypothesis confirmation, as well as to incorporate higher order evidence, such as evidence about the reliability of the sources and about its relevance with respect to the target population, hence to account both for elitists’ and pluralists’ concerns.

From a philosophical point of view, “E-Synthesis” provides a solution that accommodates the intuitions underpinning apparently conflicting concerns,[1] and has the virtue of being able to be adapted to various theoretical stances on causality (counterfactual vs. process theories vs. regularity vs. inferentialist theories of causality) (see Poellinger 2018). Section 15.5 presents how “E-Synthesis” addresses and solves philosophical issues around causal inference in medicine; in particular, the privilege accorded to randomised controlled trials, the debate around evidence hierarchies, the epistemic status of evidence for biological mechanisms, and the role of higher order evidence (evidence about evidence itself) in hypothesis confirmation.

15.2 Isolating Causes vs. Causes in Interaction: The Two Contending Paradigms

The philosophical and methodological debate concerning evidence in medicine can be roughly made sense of as a sophisticated elaboration of arguments in favour of (or against) two main paradigms for scientific inference: a categorical approach – inherited from frequentist statistics – and a probabilistic approach (where hypothesis confirmation comes in degrees) – inherited from an inductive/Bayesian approach to scientific inference.[2] Nancy Cartwright (2007a) for instance speaks about “clinching” and “vouching” methods as distinctive ways in which causal inference may be carried out in medicine and in the social sciences: vouchers, other than clinchers, do not force any conclusion, but rather suggest or support it more or less strongly.

- *Clinchers.* In principle, clinchers deductively force their conclusions and deliver an acceptance/rejection verdict on the hypothesis under investigation, on the basis of a “modus tollens” reasoning:

[1] This is possible also because the topology of our framework reveals that these conflicts relate to different levels or dimensions of causal inference and therefore can be deflated.

[2] These two approaches intersect with other epistemic stances such as empiricism vs. methodological pluralism, as well as various programs for causal inference from statistical data, however the above mentioned dichotomy can be analyzed relatively independently from these perspectives.

$$\frac{H \supset E, \neg E}{\neg H}$$

where H stays for hypothesis and E for evidence. However in experimental settings the syllogism is applied probabilistically, since the evidence $\neg E$ consists in an event that is not impossible under the assumption of H holding, but rather only very improbable. That is why $\neg E$ only invites the following statement, also called "Fisher disjunction": either something very improbable happened, but H is nevertheless true, or H is false. However, in practice hypotheses are either rejected or not, and because of this categorical approach, clinchers are characterised by a greater inductive risk, which explains why they are focused on the *probability of erroneous conclusions* (Mayo and Spanos 2006; Sprenger 2016). More importantly, the threshold for hypothesis acceptance/rejection is based explicitly on the degree of reliability that one wishes to obtain: that is, on significance levels, which are substantially based on the size of random error considered to be tolerable (for the purpose at hand). In the standard approach of hypothesis testing, the hypothesis under investigation regards the individual contribution of a putative cause to its effect (possibly moderated by some prognostic factors) and such hypothesis is contrasted with the so called null hypothesis of the cause under investigation making no difference to the observed effect. For an observed positive effect, the hypothesis space consists of three general alternatives: (1) either the treatment causes the observed effect; (2) or the effect is due to chance; (3) or the effect is due to some alternative confounding factors. Reliability theory, and methodology in general, define the instruments that should be used in order to exclude such alternative hypotheses as safely as possible: Large sample sizes help to exclude chance, by the law of large numbers (hence, the larger the sample sizes, the lower the random error, ceteris paribus); study design helps to evaluate the internal validity of the results, i.e. the extent to which alternative causes can be excluded. Therefore, clinchers are studies that are reliable both in the sense that they exclude chance (random error), and in the sense that they exclude bias or confounding (systematic error) to a large extent. In sum, clinchers maximise accuracy and internal validity.

- *Vouchers* are methods whose results cannot force any conclusion, but only suggest it. They allow only defeasible inference, but they are flexible enough to incorporate new possibly conflicting evidence in the inferential framework, without necessary eliminating old beliefs. This is a particularly appropriate approach when causes are highly context-sensitive (such in biology or the social sciences). Reliability refers here rather to whether a causal link established for some population may also hold in other contexts, and on how interacting causes modulate the causal effect.

In medical research the two approaches are reflected in two different perspectives with respect to the evaluation and use of evidence coming from different sources/methods.

The so called “Evidence Based Medicine” paradigm has emphasized the selection of “best evidence”, meant as accurate and internally valid evidence, and therefore has privileged some methods rather than others, on grounds of their higher reliability warrant. The contending view criticizes this paradigm for being uselessly and harmfully monistic and for leaning on misleading heuristics (such as evidence hierarchies).[3]

In particular, critics of the EBM approach have emphasized its shortcoming in dealing with external validity issues (Cartwright 2012; Clarke et al. 2014), which is indeed nothing else than the straightforward consequence of context sensitivity of causation, and in not paying due attention to the ontology of the phenomena being investigated (Cartwright 2007d; Anjum 2012). Furthermore, EBM is blamed for unknowingly bringing back through the window what they threw back from the door, i.e. “subjectivity” (Stegenga 2011), and for not keeping their epistemic promises (Worrall 2007b), or for running against their intended goals by not adequately distinguishing between harm vs. benefit assessment (Osimani 2014), and not being cost-effective in using the available evidence (De Pretis et al. 2019; Russo and Williamson 2007).[4]

The debate has been mainly carried on by way of illustrating case studies where one or the other approach fails to deliver the optimal information for making clinical decisions or policies. For instance, Worrall (2007b,c) illustrates the case of Extracorporeal Membrane Oxygenation for newborn babies affected by persistent pulmonary hypertension. Notwithstanding high success rate of the treatment and knowledge of the mechanisms explaining that success, efficacy was considered not to be established until proven by randomized controlled trials. This obviously exposed the newborns in the no-treatment arm to almost certain, and, given the available treatment, avoidable death. Since the newborns in the treatment arm had instead a much higher chance to survive, this was a dramatic case of lack of equipose. By pointing to the neglected sources of knowledge (mechanisms and past success record in uncontrolled case series) and by downsizing the reliability of RCTs, Worrall explains why “there is no cause to randomise” (see Worrall 2008).

Cartwright exemplifies her theory of extrapolation, by reporting about an integrated nutrition program which, while fully successful in India, completely failed in Bangladesh, and imputes this failure to a mismatch between the causal structure and social mechanisms working in the original vs. the target population (Cartwright 2012).

Also Russo and Williamson (2007) and Clarke et al. (2013, 2014) provide a rich list of case studies to support their view that evidence of different kinds are co-

[3] Also, a parallel issue concerns the strong preference accorded to the so called Potential Outcome Approach (POA) vs. more pluralistic views of causal inference and causality itself (see: Vandenbroucke et al. (2016)).

[4] Not to mention the fact that, traditionally, epistemology has considered varied evidence as more confirmatory than repetitive data (see Osimani and Landes (2020) for a discussion of the “Variety of Evidence Thesis”.)

supportive and provide a stronger basis for causal assessment, by providing distinct types of information.

Instead Howick and colleagues (2013) warn that knowledge of mechanims need always to be complemented by randomised controlled trials, while the reverse does not hold. The basis for this position is that biological pathways are extremely complicated and rich of feedback loops and backup mechanisms, which may lead to surprising outcomes at the phenotypic level. Although RCTs provide black-box evidence only, they are the only reliable basis for causal inference, in that they deliver clear information about input-output data at the level of interest, the clinical one.[5]

At its root such dissent is originated by the two sides playing different games with nature and within the scientific ecosystem. EBM plays an “elitist” game where evidence is put to test before entering the court: a certain threshold for reliability is established before the evidence is considered at all. Hence reliability is the gatekeeper and has precedence over other possible epistemic values such as the principle of total evidence, the precautionary principle, and external validity. Furthermore the notion of reliability developed with the EBM framework is method-specific; it has been constructed around a specific statistical school of thought: frequentist hypothesis testing.

Opponents of the EBM approach play a different game: they are mainly concerned by context-sensitivity of causation, and, more generally, by defeasibility of scientific inferences; consequently they appeal to the Principle of Total Evidence, as an essential desideratum in non-monotonic reasoning, as opposed to the hypothetico-deductive paradigm; furthermore, they are also worried as to the extent to which such evidence may be used for extrapolation and prediction. Hence for them also relevance of the evidence, and external validity play an important role. In both games, there is a considerable amount of uncertainty, but dissent arises as to how such uncertainty should be managed and weighted,[6] and consequently, as to the costs and utilities of each research strategy, or more generally, of any regulatory standards.[7]

[5] They cite the example of the drug “Torcetrapib”, as a case where even perfect knowledge of its functioning mechanisms did not enable its producers to avoid excess death rates in the treatment arm of the third-phase trials, and therefore denial of approval. However, this failure was in fact due to lack of knowledge about (the mechanisms leading to) the side-effect, for which knowledge of the mechanisms for the intended effect is not necessarily helpful.

[6] In the statistical counterparts of these games (frequentist vs. Bayesian statistics) probability as a measure of uncertainty is attached to the probability of error (in the long run) in one game, whereas in the other it is attached to the hypothesis itself.

[7] See also Podolsky and Powers (2015) for a critique on a recent shift in FDA evidential standards, from what I call an “elitist” to a “pluralist” view. The considerations underpinning such critiques are cast exactly in terms of the costs and benefits of each regulatory approach. Analogously Osimani (2014) and Osimani and Mignini (2015) insist that evidential standards should not be the same for assessing efficacy and harm, exactly on these epistemic and strategic grounds.

15.3 The Elitist and the Pluralist Game

The above games are associated with different epistemic goals and value specific epistemic virtues over others: it is this different preference ordering that might help us understand the dissent. Let's look a little bit closer at these preference orderings and the related payoffs and losses. In the EBM framework reliability (in both senses) is highly valued: the inferential game is seen as a game against nature, chance and biased scientists: and truth is what remains after all other factors have been eliminated. The pluralist stance is more concerned about complexity of causal phenomena and about combining a plurality of evidence to find truth in the web of interactions.

The general idea here is that whereas EBM puts efforts in decontextualising nature's signals from noise and is more focused on science distorting them (through unreliable instruments), its opponents are rather worried that nature's signals are embedded in a symphony and cannot be interpreted in isolation. Metaphorically speaking, the point is whether one considers context as noise to be eliminated, or as music which gives meaning to the individual note.

15.3.1 The Elitist Game

Clinching methods are elitist methods which value evidence for the degree to which they exclude random and systematic error. Random error refers to the "chance" variability around an otherwise reliable measurement. This can be amended by averaging the results of various replications of the study. Instead, systematic error refers to possible confounders, which systematically distort the results through replications.

The different strategies advocated by so called Evidence Based Medicine have been developed in order to secure studies from two kinds of error[8]:

1. random error;
2. systematic error (this can be generated by bias, confounding or both).

The strategies adopted to address these two kinds of error are orthogonal to each other, and are partly a consequence of the statistical approach entrenched in the EBM paradigm, that is, classical frequentist statistics.

[8] The whole enterprise of the EBM paradigm can indeed be seen as the tireless effort to systematize a set of techniques to track and possibly minimize random and systematic error.

15.3.1.1 Random Error

Random error is dealt with straightforwardly; indeed it is the basis of the entire statistical machinery of the frequentist approach: the level of tolerated *random error* is inscribed in the very methodological procedure by establishing it beforehand, and *deciding on that basis, whether the hypothesis will be accepted or rejected*, given the observed data. Hence, it is the desired "reliability level" ("significance level" and "power") that determines the ultimate fate of the hypothesis under investigation.

In this first sense, belief in the reliability of the result can be enhanced in two related ways: either by pooling results of various studies, thought to be homogeneous enough as to the sampled populations – so called meta-analyses – or by performing "identical" replications of the original study.

As much as the confidence interval may be narrow or the number of consistent replications high, these indicators of accuracy could never set you free from another kind of error, namely systematic error; that is the error arising from the in principle "identical" studies measuring the effect of a confounder instead of the investigated variable. Pure statistical considerations cannot address the problem of systematic error, which is indeed rather approached by *study design*.

It is as if we are dealing with three possible explanations when examining a given scientific result: (1) chance, (2) the investigated causal factor, (3) other causal factors (confounders, bias). Consistent replications reduce the probability of chance producing the results, but cannot do anything in discriminating between 2 and 3.

15.3.1.2 Systematic Error

The entire EBM paradigm can be seen as an effort to foster quality of evidence by both insisting on maximisation of accuracy and minimisation of systematic error (or maximisation of "internal validity"). This effort is concretised in the development of evidence hierarchies that rank studies according to their design. Hence, one finds randomised controlled trials at the top level of the ranking (and higher than that, systematic reviews and meta-analyses of RCTs); followed by comparative observational studies (such as cohort and retrospective – or, historical – studies), and below these, uncontrolled observational studies (that is, case series, and single case studies). Evidence concerning cellular mechanisms or, generally any data below the phenotypic level is ranked lowest.[9]

Internal validity refers to the neutralisation of possibly confounding factors, as candidate alternative explanations for the observed effect. These may be alternative or interactive causes (also known as prognostic and predictive factors). The preferred means to counteract this sort of problem is to use intervention and randomization, so as to avoid (self-)selection bias and have as much as possible

[9] However the reason for ranking this data below the phenotypic level lowest, is due to issues of external rather than internal validity (see Howick (2011)).

balanced groups in the experiment arms. This aims to guarantee that the observed effect is due to the treatment, and only to it. However, also random error is taken into account; hence, larger RCTs are ranked higher in the hierarchy than smaller ones, and meta-analyses of several studies are ranked higher than individual studies.

Below the standard design ranking of EBM hierarchies (with related epistemic goal) are:

1. *random vs. non-random sampling*: this criterion is important both for the *representativeness* of the sampled population (of subjects or studies) and for the external validity of the results. A randomised controlled study, whose treatment and control arms are perfectly balanced, but whose subjects have not been sampled randomly from the target population will be internally valid, but not representative of the population under study (and will suffer from low external validity if the results need to be applied to that population). Indeed samples for randomised clinical trials are almost never sampled randomly from the original population. However, systematic reviews of meta-analyses aims to achieve this same goal, in constructing a population of studies which is as representative as possible of the sampled population.
2. *experimental vs. observational design* (RCTs vs. cohort studies): experimental interventions are the best warrant of internal validity in that they severe the cause under investigation from other potential confounding factors, which could alternatively explain the observed difference in the two arms of the study;
3. *controlled vs. uncontrolled design* (cohort – or any kind of comparative studies, vs series of clinical cases with no control group of "untreated" patient). Control serves the purpose of comparing whether the same causal effect is observed also in the absence of the investigated cause; that is, to verify relevance of the putative cause to the investigated effect.

Hence, (1) control serves the goal of verifying (causal) *relevance*; (2) intervention through random allocation has the purpose of testing *causal sufficiency*; (3) and random sampling is a way to establish *causal necessity*:

1. the statistical *non-spurious* difference between outcomes in the exposed and non-exposed groups provides evidence that the presence of the putative cause makes a difference with respect to the investigated outcome;
2. random allocation ensures that no latent variable confounds the experimental results, and thus ensures that the set of causes taken into account is complete – although they are unknown. Causal sufficiency[10] is warranted by the very fact that the possible influence of all confounders (known and unknown) is neutralized through random allocation.
3. the sample population for RCTs is very rarely sampled randomly from a source population. However, the ranking of systematic reviews of meta-analyses over

[10] In the causal search literature Pearl (2000) and Spirtes et al. (2000), causal sufficiency is a fundamental assumption which grounds the algorithmic search, and undermines it if it fails to hold.

simple meta-analyses responds to the recognition that the latter may be biased by cherry-picking the included studies. Systematic reviews ensure that the sampled population *of studies* is non-biased and therefore representative of the source population. This warrants that the "causal law" inferred from the RCTs is at least valid in that specific population (the one from which the study population has been sampled from).

15.3.2 The Pluralist Game

Advocates of a pluralist methodology oppose the idea that *RCTs have a privileged role* in establishing causation (Worrall 2007a,b) as well as the view that *difference-making is necessary or sufficient* to establish causation (Cartwright 2007a). Objections to this paradigm have been raised on the following grounds:

1. *internal/external validity trade-off*: the causal structure underpinning the study results in the sample population, may not be equivalent to the one which characterises the causal dynamics in the target population, hence even studies that are strongly internally valid may fail to provide the right evidence for other kinds of populations. Indeed a sort of "inverse proportionality" relationship is posited between internal and external validity: the stricter the study protocol (inclusion/exclusion criteria, mode of administration, etc), the more likely that the study result will be internally valid, but also that its results will lead to wrong inferences in relation to real-life settings (Cartwright 2012)[11];
2. the putative privilege of randomised evidence is ill-founded (Worrall 2007b,c): that is, randomised controlled trials do not even provide any guarantee of internally valid results;

As a constructive response to these criticisms, a conciliatory position has been proposed, arguing for the adoption of a pluralistic approach to causal inference. According to this view, neither statistical evidence, nor evidence on mechanisms underpinning it are per se sufficient to establish causation, however they are both necessary. This position has been elaborated in Clarke et al. (2013, 2014), where it is advanced that different sorts of evidence may have complementary roles in supporting causal hypotheses. In particular, evidence about difference making helps de-masking causes which might be canceled out by back-up compensatory mechanisms in the organ system, whereas evidence about mechanisms is needed in order to design and interpret statistical studies – the upshot of this reasoning is that different kinds of evidence may systematically support each other and jointly (dis)confirm the causal claim under investigation.

[11] These criticisms have kindled a series of defenses of RCTs on various grounds: see Senn (2003), Papineau (1994), La et al. (2012), Teira (2011), Teira and Reiss (2013) and Osimani (2014) for an overview on the debate.

De Pretis et al. (2019) have extended this approach to various indicators of causality and relaxed the requirement that any of them be necessary, by allowing for probabilistic causal assessment, rather than categorical ones.[12]

15.3.3 Context-Sensitivity of Causality and Causal Modulation

The insistence on mechanisms and external validity from proponents of the pluralist view on medical evidence straightforwardly derives from the intuition that causation in this scientific domain is highly context sensitive. However, what does context-sensitivity exactly mean? Roughly speaking, context sensitivity refers to the joint interaction of various causes in bringing about a given effect and in modulating its intensity. Hence, subjects in studies can be conceived as vectors consisting of possible combinations of variables (at different levels: molecular, cellular, organic), that jointly contribute to the occurrence and the modulation of the effect in the presence (vs. absence) of the investigated cause.

Leonid Hanin (2017) makes this point very vividly when he explains the irreproducibility of trial results by drawing on various sources of (uncontrollable) variation in clinical research. Speaking about predictors of metastatic recurrence after breast cancer surgery, Hanin points out that:

> Whether or not a metastases will escape from dormancy in a particular patient depends not only on the effects of treatment, functioning of the immune system, concentrations of circulating angiogenesis promoters and inhibitors, and other internal factors; exhacerbation of the disease may also be triggered by intercurrent sporadic external events such as surgery unrelated to breast cancer, infection, trauma, radiation, stress, etc. Another highly significant prognostic factor is the intrinsic aggressiveness of the disease; however, its reliable assessment at early stages of the disease has proven to be far elusive. Thus the most critical determinants of the trial outcome are largely unobservable and/or unpredictable.

To make things worse, an additional level of opacity should be added to the picture (my emphasis):

> In practice the above observable prognostic factors are substituted with less informative observable surrogates such as (1) age at trial entry; (2) stage and historical grade of of the disease at surgery; (3) localization and size of the primary tumor; (4) whether or not the tumor invaded surrounding tissues; (5) the extent of nodal involvement; (6) menopausal status; (7) estrogen and progesterone receptor status; (8) presence of specific mutations in BRCA1 or BRCA2 genes; (9) family history of breast cancer; and (10) individual history of other malignancies. *Even this rough and incomplete set of surrogate clinical variables creates a large number of categories of women in both arms of the trial with potentially very different characteristics of survival. Importantly, randomization won't*

[12] This stance can be made more general by drawing on the "Variety of Evidence Thesis" (VET) which, stated in its more general form, claims that *ceteris paribus*, the more varied the evidence, the higher the confirmatory support provided to the hypothesis which explains it. Taken at face value this claim seems to favour the pluralistic methodology approach over the "evidence elitist" view adopted in the EBM paradigm. See also Osimani and Landes (2020).

> *eliminate the observable and hidden heterogeneity; it will only reduce the difference in the extent of heterogeneity between treatment and control arms.* The aforementioned inter-subject heterogeneity is quite typical of clinical trials (as opposed to in vitro experiments with cell lines or studies on animal models with tightly controlled inter-subject variation). Thus, individual responses of subjects in both arms of a trial cannot even approximately be viewed as homogeneous, let alone distributionally identical.

(Hidden) mediators and moderators may also non-additively contribute to modulating the causal effect through joint interaction. Human beings are not gas molecules; their reactions to the same treatment may be largely dependent on contingent factors (variation of individual response in different circumstances: intra-subject, or individual variation), and on various systematic mediating and interacting causes. Sample heterogeneity may be epistemically inscrutable since mediators and moderators may (and generally do) act jointly and at different organ levels (genetic/genomic, proteomic, cellular), and also emerge dynamically within one's own clinical history and social environment.

In "The Cement of the Universe" (1974) John Mackie offers a deterministic reading of so called probabilistic causation by advancing the concept of INUS condition, which provides a way to formalise the context-sensitive nature of causation. In his proposal, causes come in sets of components, (such as e.g. the presence of oxygen for a spark to start a fire), and an INUS condition is an insufficient but necessary component of an unnecessary but sufficient causal set. For instance, A is an INUS condition for E means that it is part of at least one conjunctive set standing in a biconditional relationship with E:

$$(A \wedge B) \vee (A \wedge C) \Leftrightarrow E$$

The "causal sets" can be equated to possible worlds which explain the occurrence of E in the presence of various concurring causal factors. For instance where the following holds:

$$(A \wedge B \wedge C) \vee (A \wedge D \wedge F) \vee (B \wedge D \wedge F) \Leftrightarrow E$$

the same effect E can be caused by different causal sets, such as for instance (ABC), or (ADF), or (BDF). Hence in a study which investigates the causal effect of A with respect to E (for instance a certain medical treatment with respect to a given clinical outcome) such effect will be the result of the proportion of people having also the characteristics B and C or D and F over all other subjects (since neither ACF nor ABF, nor ACD nor ABD are sufficient causal sets for E). Furthermore, there will be cases where, notwithstanding the absence of A, E will nevertheless occur (that is, when subjects present the characteristics BDF jointly).[13]

[13] Since "causal strength" is generally measured by the "effect size", that is, the proportion of subjects in the sample who show the effect E in the treatment vs. control group – for instance through relative risk ratio or odds ratio measures – causal strength can be both the result of a relative context-insensitivity of the treatment investigations (or better, the fact that it contributes

By drawing on this conceptualization of context-sensitivity, Cartwright (2012), underpins her criticisms of RCTs as gold standard in medicine, on account of the high context-sensitivity of causation in the soft sciences. Provided that the law underpinning the observed effect in relation to treatment X is formalized as follows:

$$L : Y_c = a + \beta X + W, \tag{15.1}$$

Where Y is the effect variable, X the cause variable, a is a constant, β a coefficient and W a random error summarising the effect of hidden latent variables; then, the average value of Y in the treatment group will be measured by:

$$\langle Y(u)/X(u) = x'\rangle = \langle a(u)/X(u) = x'\rangle + \tag{15.2}$$

$$\langle \beta(u)/X(u) = x'\rangle x + \tag{15.3}$$

$$\langle W(u)/X(u) = x'\rangle. \tag{15.4}$$

and, consequently, the treatment effect will be measured by the following equation:

$$T =_{def} = \langle a(u)/X(u) = x'\rangle - \langle a(u)/X(u) = x\rangle + \tag{15.5}$$

$$\langle \beta(u)/X(u) = x'\rangle x - \langle \beta(u)/X(u) = x\rangle x + \tag{15.6}$$

$$\langle W(u)/X(u) = x'\rangle - \langle W(u)/X(u) = x\rangle. \tag{15.7}$$

Since randomization is meant to warrant that the expectation of $a(u)$, $\beta(u)$ and $W(u)$ are the same whatever value X assumes (that is, that X is probabilistically independent from a, β and W), the first and last two terms in the equation cancel out; hence, the measure of the treatment effect given by an RCT results from the following formula:

$$T =_{def} \langle \beta(u)\rangle (x - x‘). \tag{15.8}$$

The critical issue raised by Cartwright is that β represents all the combinations of factors that determine not only whether X contributes to Y (if β is 0, then this contribution is null); but also how (positively or negatively) and to what extent. By drawing on the notion of causes as INUS condition, Cartwright then considers the possibility that β is a disjunction of sets of interacting factors: $\beta = f_1(z_{11}, \ldots, z_{1n}) + \ldots + f_m(z_{m1}, \ldots, z_{mp})$.

to the effect in many causal sets), or to its intrinsic force (as measured for instance by a steep dose-response curve). This intrinsic ambiguity prompted the substitution of the Bradford-Hill indicator "strength of the association" with three different indicators: probabilistic dependence, dose-response, and rate of growth in (De Pretis et al. 2019, Section 3). See Sect. 15.4 below.

This translates in an irreducible context-sensitivity component of causation, which cannot be fully accounted for by classical statistical methods of causal inference.If the main information provided by experimental evidence is about whether (and to what extent) a causal link between two phenomena exists, the information provided by detailed case series, or basic science (molecular studies; in vitro or, so called, "in silico" evidence) may contribute to a greater extent to the identification of specific subclasses in which the effect follows the treatment to a higher or lower degree, that is information about causal interactions.[14] Analogous information may also come from adjustment (stratification) and subgroup analyses in controlled studies. However, the more context-sensitive the treatment under investigation, the larger the sample sizes need to be in order to obtain such information. What is needed here is the acknowledgement that the two approaches are focused on complementary (not opposite!) goals: the "elitist approach" is concerned about establishing causal links rather independently from the task of identifying mediators and moderators, and to minimize false positives resulting from random or systematic error. Reliability of the evidence consists in the isolation of the causal link under investigation from possible interferers, and is a function of the process through which it is collected. Instead, the "pluralist approach" emphasises the irreducible context-sensitive nature of causation in medicine (and in the social sciences); therefore reliability here is about whether the acquired causal knowledge is stable enough to be applied to other contexts (prediction, external validity). Any source of information is valued as a useful basis to identify such interacting co-factors.

In the following I present a multilayer approach to causal inference in pharmacology, where both objectives are considered, and accounted for in a unifying framework.

[14] The importance to predict the variability of the effect as a function of the joint interaction of possible co-factors hence casts a new light on sources of evidence, such as case reports and case series – which standard hierarchies rank as low, because it is not "controlled" and hence cannot provide high warrant of internal validity – but whose specific epistemic import is no lower, indeed very high, in that they can provide us with valuable information about the various scenarios in which a given treatment may (or may not) induce its effect in different degrees. So, very detailed case reports facilitate inference about co-factors (prognostic factors and mediators) possibly influencing whether and to what extent, the effect size occurs in a specific population. "In silico" evidence comprehends a huge class of methodologies which can be broadly subdivided into systems biology approaches to computational modelling and simulation and machine learning techniques for knowledge extraction or pattern recognition.

15.4 E-Synthesis: A Probabilistic Causal Inference with Heterogeneous and Higher Order Evidence

The so called “reproducibility crisis” (see for instance the “Reproducibility Project: Psychology” by the Open Science Collaboration) caused some stir among psychologists, methodologists as well as scientists. The crisis extends well beyond psychology and also invests medical research (Begley and Ellis 2012; Ioannidis 2005; Prinz et al. 2011).

While some analysts have provided formal confirmation for the plausibility of such explanations (Etz and Vandekerckhove 2016; Marsman et al. 2017), and some downplay the whole issue (Senn 2002), others have further insisted on the problem of noisy data and suggested that “to resolve the replication crisis in science we may need to consider each individual study in the context of an implicit meta-analysis” (Gelman 2015). By “meta-analysis” Andrew Gelman does not mean here the standard data-pooling methods developed for instance within the Cochrane Library initiative, where statistics from sufficiently homogeneous studies are averaged so as to obtain more accurate measures of effect sizes, but rather a global approach to evidence evaluation, which takes into account not only the prima facie supportive strength any individual study gives to the investigated hypothesis (and its alternatives), but also other “higher-order” dimensions related to the entire body of evidence:

> One direction for statistical analysis that appeals to me is Bayesian inference, an approach in which data are combined with prior information *(in this case, the prior expectation that newly studied effects tend to be small, which leads us to downwardly adjust large estimated effects in light of the high probability that they could be coming largely from noise).* (Gelman 2015, p.35) (my emphasis).

Therefore, higher order evidence may consist of consistent replications across repeated measurements, knowledge about the behaviour of measurement instruments under certain conditions, or more generally, robustness of results across methods and sources, and coherence of data with the established knowledge as well as with available theories. This evidence may interact with estimation about the reliability of the source, as a function of various social factors, such as financial interests, concern for liability and regulatory issues, scientific and commercial reputation.

So far, there has been a division of labour among philosophers of science, epistemologists, and methodologists regarding these different layers. Philosophy of science and methodology (statistics) has focused on the prima facie relationship between evidence and hypothesis, and related problems (measurement error, evidential support, theory ladeness, underdetermination, internal and external validity) (Carnap 1956; Fisher 1955; Hacking 2006; Haack 2011; Lenhard 2006; Hoyningen-Huene 2013). More generally, classical and formal epistemology have paid attention to the formal warrants for knowledge justification (Audi 1993; BonJour 2009; Swinburne 2001). Social (formal) epistemology instead has concentrated its efforts on the complex framework of interests and constraints which mold the scientific ecosystem (Goldman 1999; Longino 1990; Mayo-Wilson et al. 2011).

With respect to medicine, Miriam Solomon (2015) has emphasised the social construction of medical knowledge, and David Teira has thoroughly examined the impact of the social environment (regulatory framework, economic interests, etc.) on the development of methods for evidence evaluation (see Teira, this volume, and Teira and Reiss (2013) and Teira (2011))[15]; Bennett Holman frames the evolution of regulatory tools for drug approval and monitoring as an epistemically asymmetric race of arms (Holman 2015).

In the present section I present a multilayer approach to modelling causal inference for pharmaceutical harm, that keeps track of the interactions between these different dimensions of evidence in a unitary framework (see Fig. 15.1):

1. a basic level of evidential support to the hypothesis at hand (and various evidence aggregation/amalgamation techniques, see De Pretis et al. (2019)). This constitutes the traditional focus of philosophy of science with its various approaches to scientific inference (nomological-deductive, inductive, abductive, etc.) and of statistical inference;
2. higher order level of epistemic dimensions related to the entire body of evidence: consistency/coherence of items of evidence, (in)dependence structure; reliability, relevance – this level also pertains to various meta-inferential patterns such as the "Non-Alternative-Argument" (Dawid et al. (2015)); this domain has been mainly investigated in Bayesian and formal epistemology. However, also standard statistical techniques developed to detect publication bias or other kinds of biases focus on these aspects: (Krauth et al. 2013; Rising et al. 2008; Wood et al. 2008; Lundh et al. 2017; Lundh and Bero 2017);
3. a further level comprehending information/knowledge/evidence related to such meta-epistemic dimensions themselves. This information/knowledge/evidence relates to a social epistemology level and can be more or less straightforwardly inferred from it: incentives/deterrents for bias (such as financial interests, reputation etc.), social ontology of the research domain, and nudging studies.

The philosophical foundations of such a framework have been presented in De Pretis et al. (2019). The framework consists in a Bayesian epistemic network, borrowed from Bovens and Hartmann (2003), which formalises scientific inference probabilistically. The ancestor consists in the investigated causal hypothesis, while its direct descendants are the epistemic consequences of such hypothesis, namely, indicators of causality resulting from an elaboration of Bradford-Hill guidelines for causation (Hill 1965). Concrete data is represented by reports of study results attached to the relevant causal indicator. A reliability and a relevance node input into each report node in order to weight the evidence by its level of reliability and stability with respect to the context of application (target population), see

[15] Teira analyses for instance the role played by concerns about impartiality of research in the historical establishment of randomised controlled trials as gold standards for drug approval.

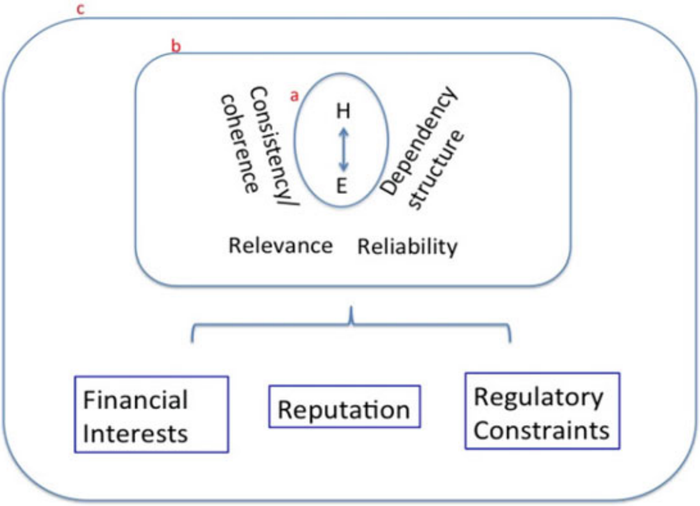

Fig. 15.1 A multilayer approach to modelling probabilistic causal inference through evidence synthesis: a. issues related to the evidential support provided by the evidence to the hypothesis at hand as discussed both in the statistical literature as well as in the philosophy of science; b. issues related to meta-evidential dimensions: consistency of studies, structure of the body of evidence in terms of mutual dependence of observations, reliability of the pieces of evidence and their relevance with respect to the target group; c. social dimensions of the pharmaceutical ecosystem (funding structures, reputational concerns, regulatory constraints), generally discussed in the social epistemology literature

Fig. 15.2. By breaking down the evidential line between pieces of evidence and causality into a two-stage process mediated by causal indicators, E-Synthesis helps disentangle philosophical issues related to the conceptualisation of causality from those related to causal inference and diagnostics.[16] At least some of the disputes among philosophers and methodologists may be solved by keeping these levels as distinct (see 15.5).

Figure 15.3 illustrates the structure of the causal inference problem: the set of dependencies and independencies determines a hierarchical structure where causal indicators – difference making (Δ), probabilistic dependency (PD), dose-response relationship (DR), rate of growth (RoG), evidence of mechanism (M), information about time asymmetries (T), etc.[17] lie on the same level, whereas study reports, possibly deriving from different methods (observational, such as cohort or

[16] This is also in analogy with Bogen and Woodward's distinction between data and phenomena (Bogen and Woodward 1988).

[17] These are derived from the epidemiological/causal literature (in particular from Bradford Hill guidelines); see De Pretis et al. (2019).

retrospective studies) or experimental (such as Randomised Controlled Studies) lie below this level.[18]

To each report a reliability node is attached: this is intended to capture the degree of systematic error (confounding and bias) estimated to affect the source report, as well as a relevance node, referring to the representativeness of the study sample for the purpose of causal inference with respect to the target population. Random error is directly represented by the likelihood function mapping reports to abstract indicators.

The graphical form provides an illustration of the epistemic dimensions at stake and thereby provides greater insight into some methodological issues by offering a mathematical explanation of their dynamics. This sort of representation allows one to single out in the mathematical formulae the specific role played by each epistemic dimension in the inferential dynamics; e.g., the role of reliability with respect to the propagation of confirmation in connection with replication of studies and with heterogenous sources of evidence. This framework has several advantages[19]:

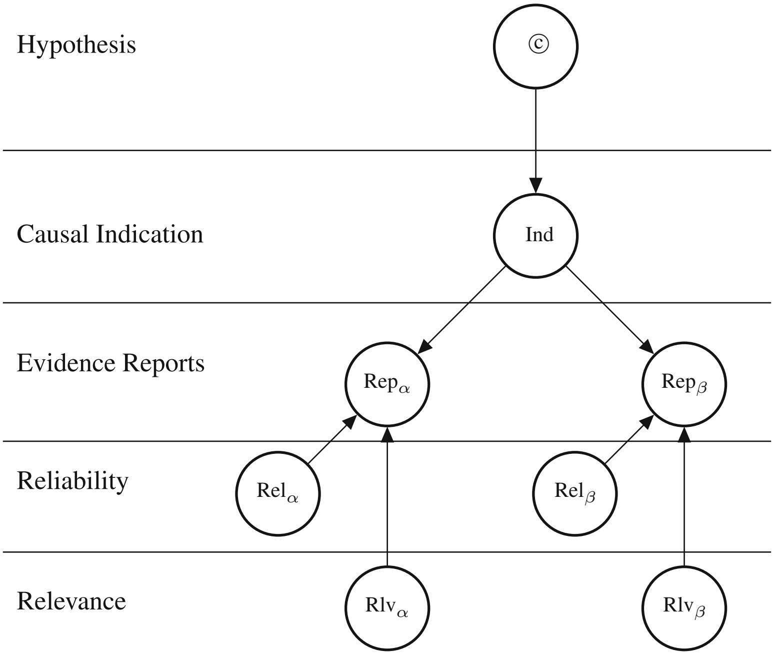

Fig. 15.2 Graph structure of the Bayesian network for two reports and epistemic categories

[18] The frameworks also allows modeling and simulation studies to play a relevant confirmatory role, especially with regard to the underlying dynamics underpinning the phenotypic causal effect. See also Osimani and Poellinger (2020).

[19] Please refer to De Pretis et al. (2019) for basics and details of the framework.

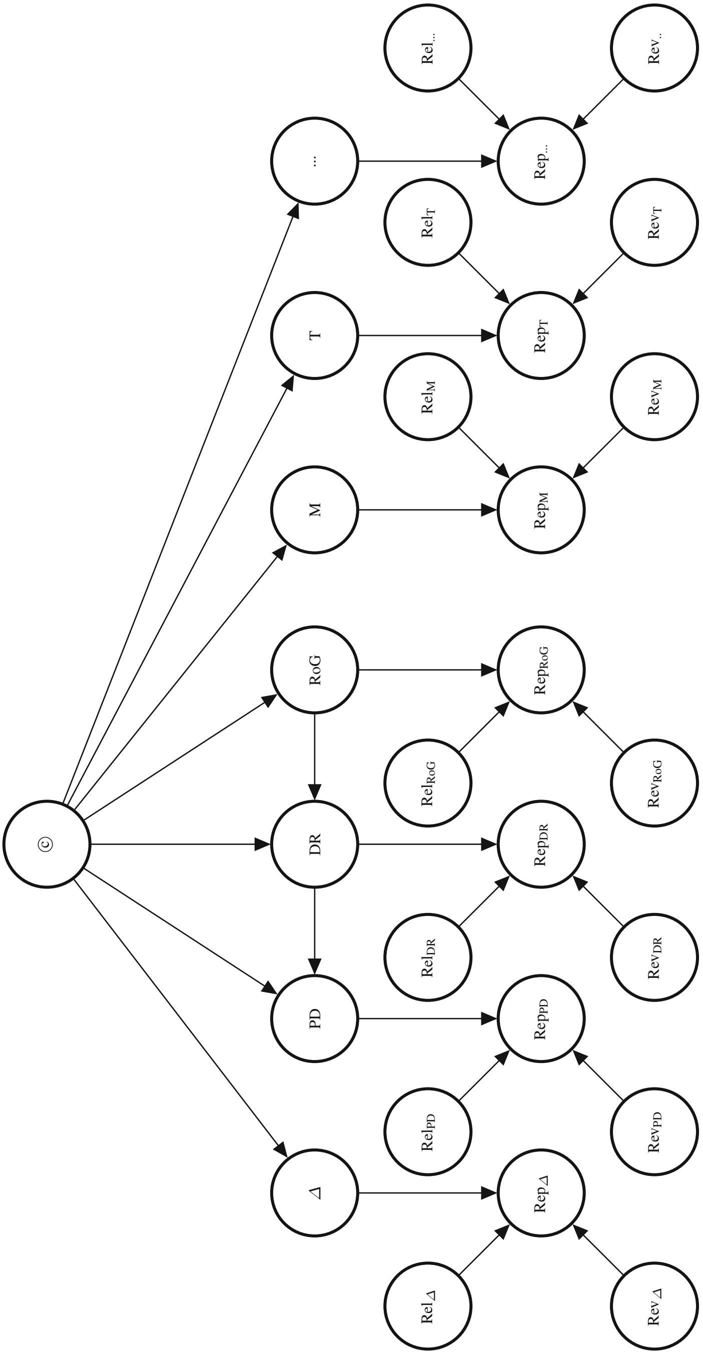

Fig. 15.3 Graph of the Bayesian network with one different report for every causal indicator variable. The dots indicate that there might be further indicators of causality not considered here

1. it identifies possible indicators of causality on the basis of the methodological and philosophical literature on causality, evidence, and causal inference;
2. embeds them in a topological framework of probabilistic dependencies and independencies grounded in assumptions regarding their reciprocal epistemic interconnections;
3. weakly orders some of these probabilistic dependencies as a function of their inferential strength with respect to the confirmation of causal hypotheses.
4. it easily accommodates many intuitions already expressed by philosophers of medicine regarding pluralistic approaches to evidence evaluation;
5. it lends itself to explicitly track the interaction of several dimensions of evidence, such as coherence and reliability;
6. it allows for a pluralistic but at the same time systematic approach to evidence amalgamation;
7. it tracks the different role of cross-validation through heterogeneous methods or sources vs. the confirmatory contribution of exact replication (see also Osimani and Landes (2020));

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

15.5 Discussion

By breaking down the evidential line between pieces of evidence and causality into a two-stage process mediated by causal indicators, E-Synthesis helps disentangle philosophical issues related to the conceptualisation of causality from those related to causal inference and diagnostics. Furthermore, the framework aims to probabilistic causal assessment, and therefore bypasses problems generated by accounts that aim to establish causal claims categorically (through the identification of necessary and sufficient conditions for causation). More importantly, separate nodes for relevance and reliability are embedded in the network; this contributes to deflate much discussion in the methodological literature concerning the trade-off between these two dimensions of evidence quality.

15.5.1 The EBM vs. Pluralist Approach to Causal Inference

Translated within the E-Synthesis framework, the implicit assumptions underpinning the EBM viewpoint is that ideal RCTs (that is internally valid ones) are perfect indicators of difference making, and difference making is a perfect indicator of causality,[20] whereas other indicators only weakly support the hypothesis of causation. This can be represented in logical terms as an entailment relationship:

$$RCT \supset \Delta \supset \copyright.$$

As a consequence, EBM focuses on the level of evidence that goes through the Δ indicator, and concentrates its efforts on having as reliable as possible evidence for such indicator.

The contending view is that different indicators may have complementary epistemic roles in supporting the hypothesis of causality. However, to this view, held for instance by Clarke et al. (2014), Howick counters that: “There are many cases where patient-relevant effects of medical therapies have been established by comparative clinical studies alone.” (Howick 2011, p. 939).

The debate is jeopardised by conflating the two entailment relationships RCT $\supset \Delta$, and $\Delta \supset \copyright$ into one; that is, what is discussed is whether it can be justifiably held that RCTs provide perfect information for causality (directly): i.e. whether

$$RCT \supset \copyright,$$

rather than either whether RCT $\supset \Delta$, or $\Delta \supset \copyright$.

If we let the epistemic net of E-Synthesis represent this discussion, we can take Howick to hold the view that:

$$P(\copyright \mid \Delta) = 1,$$

and that ideal RCTs provide strong evidence for Δ:

$$P(\Delta \mid RCT) \approx 1,$$

hence offering strong enough evidence to establish the causal claim; while for the contenders $P(\copyright \mid \Delta)$ is too small to establish the causal claim. For them, also having mechanistic evidence is required to establish the causal claim:

$$P(\copyright \mid \Delta \& M) = 1.$$

E-Synthesis allows for RCT $\supset \Delta \supset \copyright$ to hold, but the entailment relationship between RCTs and Δ is one between ideal RCTs and Δ. Hence in any concrete case $P(\Delta|\text{RCT}) < 1$, and this leaves room for other kinds of evidence to also contribute to hypothesis confirmation. On the other side, by dropping any necessity

[20] This directly derives from the potential outcome approach underpinning RCT methodology. See Holland et al. (1985), Rubin (2005), and Vandenbroucke et al. (2016) for a critical appraisal of this approach.

or sufficiency requirements for causal inference, this approach relaxes the theory that both evidence of association and of underlying biological mechanisms is necessary to establish causation. This approach is therefore much more flexible and responds both to the EBM intuition underpinning the privileged role assigned to RCTs, as well as to the pluralist intuition that various kinds of evidence are contributory to causal assessment.

15.5.2 Evidence Hierarchies

In relation to evidence hierarchies – which is a strong point of contention among philosophers of medicine and methodologists – E-Synthesis poses a set of inequalities regarding the evidential strength of various causal indicators (see De Pretis et al. (2019, p.32–33)), which nicely parallels the rankings proposed in the EBM paradigm:

$$P(©|\Delta) > P(©|DR), P(©|RoG) > P(©|PD) \tag{15.9}$$

What differentiates E-Synthesis from standard evidence rankings however is that these have predominantly been formalised as lexicographic decision rules. This means that higher-level studies trump lower-level ones: when two studies of different levels deliver contradictory findings, then the one higher in the evidence hierarchy is considered more reliable and allows one to discard the lower level one.[21] E-Synthesis incapsulates the rationale for ranking evidence (in the inequalities across probabilistic dependencies between causation and various indicators), but at the same time allows one to take into account all evidence, and to act accordingly, as soon as the probability of the causal hypothesis goes above the threshold established by the other dimensions of the decision (utility of withdrawing/not withdrawing the drug, conditional on the probability of it causing the suspected harm) (see De Pretis et al. 2019, Section 2.2.).[22]

15.5.3 Causal Holism

Methodological pluralists such as Cartwright (2011; 2007b), and Stegenga (2011), among others, express concerns against the privileged role of RCTs also on grounds

[21] A somewhat unwanted consequence of this "take the best" approach is that it has become commonplace to assume an uncommitted attitude towards observed associations least they are "proved" by gold standard evidence (see the still ongoing debate on the possible causal association between paracetamol and asthma: Osimani (2014)).

[22] This also complies with the precautionary principle in risk assessment and with how decisions should be made in health settings: Osimani (2007, 2013) and Osimani et al. (2011).

that classical ‘linear’ approaches to causal inference cannot do justice to the complexity of causal phenomena in the biological and social sciences, characterized by nonlinear causation and causal interactions.[23]

Strictly speaking this sort of criticism does not deny that:

$$\Delta \supset ©,$$

but only denies the reverse:

$$© \supset \Delta.$$

Since in the causal graph literature the defining features for causality jointly entail that $\Delta \Leftrightarrow ©$ but not the reverse, this criticism misses the point. We contribute to deflate the debate by not collapsing Δ and © into a single node, therefore allowing causation to be holistic and therefore not reducible to difference-making, and at the same time letting difference-making immediately imply causation: in E-Synthesis when a difference-making relationship between two events or variables holds, then this is a sufficient – although not necessary – condition for causality. This can be characterized in logical terms as an entailment relationship: $\Delta \supset ©$. Hence, in E-Synthesis, the probability of a causal relationship, given a genuine difference making relationship is 1: $P(© \mid \Delta) = 1$. The inverse entailment though, $© \supset \Delta$, does not hold: knowledge of © does not necessitate the existence difference-making – e.g., in cases of “holistic causation”.[24]

15.5.4 *Evidence of Mechanisms and Relevance*

Pluralists accord to evidence of mechanisms a preeminent role in establishing external validity and extrapolation, in that it helps evaluate whether the cause under investigation will work in a similar “context” also in the target population (Russo and Williamson 2007; Cartwright 2007a; Cartwright and Stegenga 2011). The present framework however, formally distinguishes the role of evidence of mechanisms for the purpose of causal assessment, from its role for the purpose of establishing external validity by associating the latter to the relevance node RLV. This allows us to explicitly distinguish the different kinds of inductive risk involved in the inference: (1) from statistical dirty data to causal indicator(s), and then to causality, in a specific population, (2) the extrapolation of a causal link established

[23] In the same line, also modular conceptualization of causes such as the ones implied in the causal graph methodology developed by Pearl (2000) and Glymour Spirtes et al. (2000) and colleagues (see also Woodward (2003)), are under attack for failing to recognize that causes may be holistic and therefore may be not adequately captured by a difference making account.

[24] This responds to concerns expressed among others by Cartwright (2007c), Mumford and Anjum (2011), Anjum (2012), and Kerry et al. (2012).

in a given population/model to another population/model. More importantly, adding a relevance node to the evidence reports allow for even highly reliable RCTs to play a low evidential support if they are not considered to be relevant to the target population.

15.5.5 Reliability and Higher Order Evidence

By introducing a reliability node *Rel*, and thereby breaking up the different dimensions of evidence (strength, relevance, reliability) E-Synthesis allows them to be explicitly tracked in the body of evidence. This makes it possible to parcel out the strength of evidence from the method with which it was obtained.[25] With this, E-synthesis provides a higher order perspective on evidential support by effectively embedding these various epistemic dimensions in a concrete topology. Indeed, the framework presented here also provides a fruitful platform for integrating insights developed in the philosophy of science around such topics as the role of replication in assessing the reliability of evidence (Open Science Collaboration 2015; Meehl 1990; Lamal 1990; Hempel 1968; Platt 1964), as well as the confirmatory role of explanatory power (McGrew 2003; Crupi et al. 2013; Cohen 2016; Lipton 2003) and coherence (Dietrich and Moretti 2005; Moretti 2007; Wheeler and Scheines 2013; Fitelson 2003; Bovens and Hartmann 2003). This paper focuses on inference *within one model*, rooting in *one hypothesis*, but E-Synthesis allows for going beyond the network's limits and for embedding it in an even larger network to trace the hypothesis' relation with other potentially concurring hypotheses. The mechanics of Bayesian epistemology are flexible enough to permit such an augmentation for the purposes of tracing further inference patterns. The framework is currently being developed into a concrete tool for evidence amalgamation (see Landes et al. (2018)) and possibly into a software.

Acknowledgements This research was supported by the European Research Council, ERC Starting Grant GA 639276: "Philosophy of Pharmacology: Safety, Statistical Standards, and Evidence Amalgamation". I thank my colleagues from the Munich Center for Mathematical Philosophy, and various audiences to which different parts of this talk were presented:

I also thank previous colleagues and students at the MCMP who provided precious comments and feedback on earlier versions of the paper. I am also grateful for discussions on the topics mentioned in the paper with Rani Anjum, Jeffrey Aronson, Bengt Autzen, Seamus Bradley, Lorenzo Casini, Vincenzo Crupi, Ralph Edwards, Branden Fitelson, Stephan Hartmann, Erik Nyberg, Elena Rocca, Jan Sprenger, Daniel Steel, and Momme von Sydow. Finally, I thank Adam LaCaze, Scott Podolsky, David Teira for reading and commenting on earlier drafts of the paper, with the usual disclaimer that any errors are my own responsibility.

[25] Osimani and Landes (2020) investigates the various concepts of reliability involved in such considerations.

References

Anjum, R. L., & Mumford, S. (2012). Causal dispositionalism, Chap. 7. In Bird, A., Ellis, B., Sankey, H. (Eds.), *Properties, powers and structure* (pp. 101–118). New York: Routledge.

Audi, R. (1993). *The structure of justification*. Cambridge: Cambridge University Press.

Beauchamp, T. L. (2011). Informed consent: Its history meaning, and present challenges. *Cambridge Quarterly of Healthcare Ethics, 20*(04), 515–523.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*(7391), 531–533.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The Philosophical Review, 97*(3), 303–352.

BonJour, L. (2009). *Epistemology: Classic problems and contemporary responses*. Lanham: Rowman & Littlefield Publishers.

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

Carnap, R. (1956). The methodological character of theoretical concepts. Indianapolis: Bobbs-Merrill.

Cartwright, N. (2007a). Are RCTs the gold standard? *Biosocieties, 2*, 11–20. https://doi.org/10.1017/S1745855207005029.

Cartwright, N. (2007b). Are RCTs the gold standard? *BioSocieties, 2*(1), 11–20. https://doi.org/10.1017/S1745855207005029.

Cartwright, N. (2007c). *Causal powers: What are they? Why do we need them? What can be done with them and what cannot?* Technical report, contingency and dissent in science technical report 04/07. http://www.lse.ac.uk/CPNSS/research/concludedResearchProjects/ContingencyDissentInScience/DP/CausalPowersMonographCartwrightPrint%20Numbers%20Corrected.pdf.

Cartwright, N. (2007d). *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge/New York: Cambridge University Press.

Cartwright, N. (2012). Presidential address: Will this policy work for you? Predicting effectiveness better: How philosophy helps. *Philosophy of Science, 79*(5), 973–989. https://doi.org/10.1086/668041.

Cartwright, N., & Stegenga, J. (2011). A theory of evidence for evidence-based policy. In Dawid, P., Twining, W., Vasilaki, M. (Eds.), *Evidence inference and enquiry* (Chapter 11, pp. 291–322). Oxford: Oxford University Press.

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2013). The evidence that evidence-based medicine omits. *Preventive Medicine, 57*(6), 745–747. https://doi.org/10.1016/j.ypmed.2012.10.020.

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2014). Mechanisms and the evidence hierarchy. *Topoi, 33*, 339–360. https://doi.org/10.1007/s11245-013-9220-9.

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

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

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

De Pretis, F., Landes, J., & Osimani, B. (2019). E-synthesis: A Bayesian framework for causal assessment in Pharmacosurveilance. *Accepted in Frontiers in Pharmacology*.

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

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS One, 11*(2), e0149794.

Faden, R. R., & Beauchamp, T. L. (1986). *A history and theory of informed consent*. New York: Oxford University Press.

Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society Series B (Methodological), 17*, 69–78.

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

Gelman, A. (2015). Working through some issues. *Significance, 12*(3), 33–35. https://doi.org/10.1111/j.1740-9713.2015.00828.x.

Goldman, A. I. (1999). *Knowledge in a social world* (Vol. 281). Oxford/New York: Clarendon Press Oxford.

Haack, S. (2011). *Defending science-within reason: Between scientism and cynicism*. New York: Prometheus Books.

Hacking, I. (2006). *The emergence of probability: A philosophical study of early ideas about probability induction and statistical inference*. Cambridge: Cambridge University Press.

Hanin, L. (2017). Why statistical inference from clinical trials is likely to generate false and irreproducible results. *BMC Medical Research Methodology, 17*(1), 127. https://doi.org/10.1186/s12874-017-0399-0.

Hempel, C. G. (1968). Maximal specificity and lawlikeness in probabilistic explanation. *Philosophy of Science, 35*(2), 116–133. http://www.jstor.org/stable/186482.

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

Holland, P. W., Glymour, C., & Granger, C. (1985). Statistics and causal inference. *ETS Research Report Series, 1985*(2), i–72.

Holman, B. (2015). *The fundamental antagonism: Science and commerce in medical epistemology.* PhD Dissertation. Irvine: University of California.

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

Howick, J., Glasziou, P., & Aronson, J. K. (2013). Problems with using mechanisms to solve the problem of extrapolation. *Theoretical Medicine and Bioethics, 34*(4), 275–291.

Hoyningen-Huene, P. (2013). *Systematicity: The nature of science.* New York: Oxford University Press.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124.

Joffe, M. (2011). Causality and evidence discovery in epidemiology. In D. Dieks, W. J. Gonzalez, S. Hartmann, T. Uebel, & M. Weber (Eds.), *Explanation, prediction, and confirmation* (pp. 153–166). Dordrecht: Springer Netherlands. ISBN: 978-94-007-1180-8. https://doi.org/1.1007/978-94-071180-8_11.

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

Krauth, D., Woodruff, T. J., & Bero, L. (2013). Instruments for assessing risk of bias and other methodological criteria of published animal studies: A systematic review. *Environmental Health Perspectives, 121*(9), 985.

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

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

Landes, J., Osimani, B., & Poellinger, R. (2018). Epistemology of causal inference in pharmacology. *European Journal for Philosophy of Science, 8*(1), 3–49.

Lenhard, J. (2006). Models and statistical inference: The controversy between Fisher and Neyman–Pearson. *The British Journal for the Philosophy of Science, 57*(1), 69–91.

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

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton: Princeton University Press.

Lundh, A., & Bero, L. (2017). The ties that bind. *British Medical Journal, 356*. https://doi.org/10.1136/bmj.j176.

Lundh, A., Lexchin, J., Mintzes, B., Schroll, J. B., & Bero, L. (2017). Industry sponsorship and research outcome. *Cochrane Library, 2*, Art. No.: MR000033. https://doi.org/10.1002/14651858.MR000033.pub3.

Marsman, M., Schoönbrodt, F. D., Morey, R. D., Yao, Y., Gelman, A., & Wagenmakers, E.-J. (2017). A Bayesian bird's eye view of 'Replications of important results in social psychology'. *Royal Society Open Science, 4*(1), 160426.

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *British Journal for the Philosophy of Science, 57*(2), 323–357. https://doi.org/10.1093/bjps/axl003.

Mayo-Wilson, C., Zollman, K. J. S., & Danks, D. (2011). The independence thesis: When individual and social epistemology diverge. *Philosophy of Science, 78*(4), 653–677.

McGrew, T. (2003). Confirmation, heuristics, and explanatory reasoning. *British Journal for the Philosophy of Science, 54*(4), 553–567. http://www.jstor.org/stable/3541678.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102%5C_1.

Moretti, L. (2007). Ways in which coherence is confirmation conducive. *Synthese, 157*(3), 309–319. https://doi.org/10.1007/s11229-006-90575

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716.

Osimani, B. (2007). *Probabilistic information and decision making in the health context. The package leaflet as a basis for informed consent* (1st edn). Lugano: USI, Università della Svizzera italiana. http://doc.rero.ch/record/28759?ln=fr.

Osimani, B. (2012). Risk information processing and rational ignoring in the health context. *The Journal of Socio-Economics, 41*(2), 169–179.

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

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

Osimani, B., & Mignini, F. (2015). Causal assessment of pharmaceutical treatments: Why standards of evidence should not be the same for benefits and harms? *Drug Safety, 38*(1), 1–11. ISSN: 1179–1942. https://doi.org/10.1007/s40264-014-0249-5.

Osimani, B. & Landes, J. (2020). *Varieties of Error and Varieties of Evidence in Scientific Inference*. (Accepted).

Osimani, B., & Poellinger, R. (2020). A protocol for model validation and causal inference from computer simulation. In M. Bertolaso & F. Sterpetti (Eds.), *A critical reflection on automated science. Will science remain human*. Heidelberg: Springer Nature. (forthcoming).

Osimani, B., Russo, F., & Williamson, J. (2011). Scientific evidence and the law: An objective Bayesian formalization of the precautionary principle in pharmaceutical regulation. *The Journal of Philosophy Science & Law, 11*(2), 1–24.

Papa, A. (2014). *L'identità esposta. La cura come questione filosofica.* Milano: Vita e Pensiero.

Papineau, D. (1994). The virtues of randomization. *The British Journal for the Philosophy of Science, 45*(2), 437–450.

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

Pessina, A. (2009). Biopolitica e Persona. *Medicina e Morale, 2*, 239–253. http://hdl.handle.net/10807/4748.

Platt, J. R. (1964). Strong inference. *Science, 146*(3642), 347–353.

Podolsky, S. H., & Powers, J. H. (2015). Regulating antibiotics in an era of resistance: The historical basis and continued need for adequate and well-controlled investigations regulating antibiotics in an era of resistance. *Annals of Internal Medicine, 163*(5), 386–388.

Poellinger, R. (2018). On the ramifications of theory choice in causal assessment. Indicators of causation and their conceptual relationships. Submitted.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery, 10*(9), 712–712.

Rising, K., Bacchetti, P., & Bero, L. (2008). Reporting bias in drug trials submitted to the Food and Drug Administration: Review of publication and presentation. *PLoS Medicine, 5*(11), e217.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association, 100*, 322–331.

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

Scheu, G. (2003). *In dubio pro securitate: Contergan, Hepatitis-/AIDS-Blutprodukte, Spongiformer Humaner Wahn und kein Ende?* (Vol. 42). Baden-Baden: Nomos.

Senn, S. (2002). A comment on replication, p-values and evidence SN Good-man, statistics in medicine 1992; 11: 875–879. *Statistics in Medicine, 21*(16), 2437–2444.

Senn, S. (2003). *Dicing with death: Chance risk and health.* Cambridge: Cambridge University Press.

Sgreccia, E. (2007). *Manuale di Bioetica* (3rd ed.). Milano: Vita e Pensiero.

Solomon, M. (2015). *Making medical knowledge.* Oxford: Oxford University Press.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search.* Cambridge: MIT press.

Sprenger, J. (2016). Bayesianism vs. Frequentism in statistical inference (Chap. 18). Oxford University Press.

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

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

Swinburne, R. (2001). *Epistemic justification*. Oxford: Oxford University Press.

Teira, D. (2011). Frequentist versus Bayesian clinical trials. In Gifford, F. (Ed.), *Handbook of philosophy of medicine* (pp. 255–298). Amsterdam: Elsevier.

Teira, D., & Reiss, J. (2013). Causality, impartiality and evidence-based policy. In H.-K. Chao, S.-T. Chen, & R. L. Millstein (Eds.), *Mechanism and causality in biology and economics* (pp. 207–224). Dordrecht: Springer. https://doi.org/10.1007/978-94-007-2454-9_11.

Vandenbroucke, J. P., Broadbent, A., & Pearce, N. (2016). Causality and causal inference in epidemiology: The need for a pluralistic approach. *International Journal of Epidemiology, 45*, 1776–1786. https://doi.org/10.1093/ije/dyv341.

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

Wood, L., Egger, M., Gluud, L. L., Schulz, K. F., Juüni, P., Altman, D. G., Gluud, C., Martin, R. M., Wood, A. J. G., & Sterne, J. A. C. (2008). Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: Meta-epidemiological study. *BMJ, 336*(7644), 601–605.

Woodward, J. (2003). *Making things happen: A theory of causal explanation (Oxford studies in the philosophy of science).* New York: Oxford University Press. ISBN: 9780195189537.

Worrall, J. (2007a). Do we need some large, simple randomized trials in medicine? *EPSA Philosophical issues in Science*, 289–301. https://doi.org/10.1007/9789048132522_27.

Worrall, J. (2007b). Evidence in medicine and evidence-based medicine. *Philosophy Compass, 2*(6), 981–1022. https://doi.org/10.1111/j1747.9991.2007.00106.x.

Worrall, J. (2007c). Why there's no cause to randomize. *British Journal for the Philosophy of Science, 58*(3), 451–88. https://doi.org/10.1093/.bjps/axm024.

Worrall, J. (2008). Evidence and ethics in medicine. *Perspectives in Biology and Medicine, 51*(3), 418–431. https://doi.org/10.1353/pbm.0.0040.