Synthese manuscript No.
(will be inserted by the editor)

Evidence Amalgamation in the Sciences: An Introduction

Samuel C. Fletcher · Jürgen Landes · Roland Poellinger

Received: date / Accepted: date

Abstract Amalgamating evidence from heterogeneous sources and across levels of inquiry is becoming increasingly important in many pure and applied sciences. This special issue provides a forum for researchers from diverse scientific and philosophical perspectives to discuss evidence amalgamation, its methodologies, its history, its pitfalls and its potential. We situate the contributions therein within six themes from the broad literature on this subject: the variety-of-evidence thesis, the philosophy of meta-analysis, the role of robustness/sensitivity analysis for evidence amalgamation, its bearing on questions of extrapolation and external validity of experiments, its connection with theory development, and its interface with causal inference, especially regarding causal theories of cancer.

Keywords Variety-of-evidence thesis · Meta-analysis · Sensitivity analysis · Extrapolation · External validity · Theory development · Causal inference · Causal theories of cancer

1 Introduction

The amalgamation of evidence from different models, scales, and types of data continues to be central in diverse sciences such as biology, ecology, medicine,

S. C. Fletcher
University of Minnesota, 831 Heller Hall, 271 19th Ave S, Minneapolis, MN 55455, USA
E-mail: scfletch@umn.edu

J. Landes
LMU Munich, Geschwister-Scholl-Platz 1, 80539 Munich, Germany
E-mail: Juergen.Landes@lrz.uni-muenchen.de

R. Poellinger
Municipal Department of Arts and Culture of the City of Munich, Rosenheimer Straße 5, 81667 Munich, Germany
E-mail: Roland.Poellinger@outlook.com

sociology, geography, climate science and economics. When access to phenomena of interest is incomplete, piecemeal, indirect, or mediated by substantial auxiliary assumptions, it is not always obvious in what manner scientists can justifiably decide how their total evidence comparatively supports hypotheses and informs future research. Policy makers, professional practitioners, and others must act appropriately informed by such complex and heterogeneous evidence. And philosophers of science try to understand the underlying logic of these practices, their role in the history and development of the sciences, and their avenues for refinement. Accordingly, the critical analysis of evidence amalgamation in the sciences involves historical and descriptive aspects as well as epistemically, methodologically, and ethically normative ones.

Here we have gathered thirteen contributions along each of these lines of inquiry. With such diversity, it is of course difficult to amalgamate their collective morals into a simple conclusion! So, in this introduction, we introduce the basic concepts and questions within the main themes of this special issue (in section 2) before situating the contributions to the issue within these themes and (in section 3) describing them in more detail. For readers interested in specific themes, here is a list thereof with corresponding contributions:

1. The variety-of-evidence thesis (section 2.1): Claveau and Grenier (section 3.1) and Heesen et al (section 3.2).
2. The philosophy of meta-analysis (section 2.2): Holman (section 3.3), Vieland and Chang (section 3.4), and Wüthrich and Steele (section 3.5).
3. The role of robustness/sensitivity analysis in amalgamating diverse evidence (section 2.3): Wüthrich and Steele (section 3.5), Wilde and Parkkinen (section 3.6), and Kao (section 3.10).
4. How diverse types of evidence bear on the external validity of experimental conclusions and extrapolation therefrom (section 2.4): Wilde and Parkkinen (section 3.6), Frank (section 3.7) and Reiss (section 3.8).
5. The role of amalgamating diverse evidence in theory development (section 2.5): Bertolaso and Sterpetti (section 3.9) and Kao (section 3.10).
6. Causal inference from diverse evidence (section 2.6): Wilde and Parkkinen (section 3.6), Danks and Plis (section 3.11), Mayo-Wilson (section 3.12), and Baetu (section 3.13).

Among these, Wilde and Parkkinen (section 3.6), Reiss (section 3.8) and Bertolaso and Sterpetti (section 3.9) use the example of causal theories of cancer to illustrate their arguments, hence are of interest to readers topically interested in how those theories meet general issues in amalgamating evidence.

Finally, in section 4, we describe our outlook on these themes, including future directions for research, in light of these contributions.

2 Diverse Topics

2.1 Variety of Evidence

Varied evidence for a hypothesis confirms it more strongly than less varied evidence, ceteris paribus. This epistemological Variety of Evidence Thesis enjoys widespread and long-standing intuitive support among scientific methodologists (Carnap, 1962; Earman, 1992; Horwich, 1982; Hüffmeier et al, 2016; Keynes, 1921).

Nowadays, confirmation is almost always understood in terms of a Bayesian confirmation measure. The starting point of contemporary Bayesian analyses of this thesis (Claveau, 2013; Landes, 2018; Landes and Osimani, 2019; Stegenga and Menon, 2017) is the analysis of Bovens and Hartmann (2002, 2003)—the recent Kuorikoski and Marchionni (2016) is an interesting exception to this rule.

Bovens and Hartmann study the Bayesian confirmation a body of evidence, $\mathcal{E}$, bestows on a (scientific) hypothesis of interest, H. Their model of scientific inference is represented by a Bayesian network over the following binary variables: a hypothesis variable H, a set of variables C which represent the testable consequences of the hypothesis H, evidence variables E, each of which pertain to precisely one consequence variable, and a set of variables R which stand for the reliability of the instruments employed to obtain the evidence. To compare two different bodies of evidence $\mathcal{E}_D$ and $\mathcal{E}_N$ in terms of their confirmatory value, they compare the difference in posterior beliefs in the hypothesis H, i.e., they compare $P(H|\mathcal{E}_D)$ to $P(H|\mathcal{E}_N)$.

They compare the confirmation of an hypothesis H by a diverse body of evidence (depicted in the right-hand column of Figure 1), $\mathcal{E}_D$, to that by a narrow body of evidence (depicted in the left-hand column of Figure 1), $\mathcal{E}_N$, in three different scenarios. Plausibly, one may take the Variety of Evidence Thesis to entail in all three scenarios that

$$P(H|\mathcal{E}_D) > P(H|\mathcal{E}_N) \ .$$

Their key contribution is to show that instead

$$P(H|\mathcal{E}_D) < P(H|\mathcal{E}_N)$$

in all three scenarios for some sensible prior probability distributions.

Claveau and Grenier (2018) extend the Bovens and Hartmann model by formalizing the notion of unreliability in a way which is closer to scientific practice. Furthermore, they consider consequence variables and reliability variables that are dependent to a degree, showing that the VET fails in many of their models. Heesen et al (2018) break with the tradition of a Bayesian analysis of the confirmatory value of varied evidence obtained by employing diverse methods. In their contribution drawing on voting theory, it is the evidence from diverse methods which supports the hypothesis of interest more strongly.

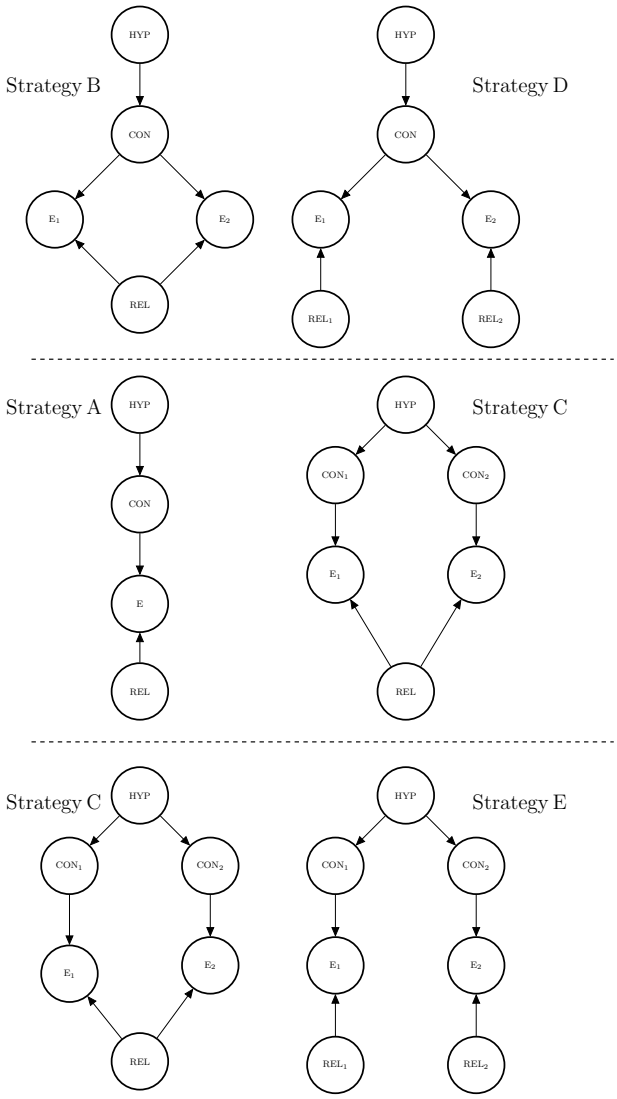

Fig. 1 The three scenarios described in Bovens and Hartmann (2003) as depicted in Landes and Osimani (2019): each row represents a scenario comparing two parallel strategies: B vs. D in the upper row, A vs. C in the middle, and C vs. E in the lowest row.

2.2 Meta-Analysis

Meta-analysis is the branch of statistics concerned with how to amalgamate evidence for hypotheses from many different individual studies, each typically with their own analysis. A meta-analysis is, as it were, an analysis of analyses, designed to produce a summary report on the evidence from varied sources.

Even when scientific studies designed to measure an effect within a target population study are well-designed, their conclusions are still defeasible at least because of the random variation within that population. In other words, random variation in the data actually sampled can entail misleading conclusions either for or against hypotheses and variation in estimated sizes of effects of interest. But various asymptotic results in statistics provide some assurance that the probability of being so mislead becomes smaller and smaller as more data are collected—as the evidence accumulated from different statistical studies is assessed together, instead of separately. Thus procedures for meta-analysis are procedures for the amalgamation and assessment of (potentially) the total statistical evidence available for hypotheses and effects.

The statistical and scientific literature on the technical aspects of how to perform a meta-analysis is huge—see, e.g., Sutton et al (2001); Sutton and Higgins (2008); Cumming (2012) for reviews—in contrast with the near complete absence of discussion of those technical aspects' conceptual and epistemological foundations among philosophers of science—but see Kilinc (2012) and Vieland and Chang (2018), the latter of whom raise interesting puzzles about their epistemic justification. Rather, philosophical attention to meta-analysis has so far focused on the role of the social structure of science in assessing the cogency and objectivity of meta-analytic procedures, especially for biomedical research. This is due to the explicit recognition that the Evidence-Based Medicine (EBM) community gives to meta-analysis of randomized controlled trials (RCTs) at the top of proposed evidential hierarchies (Reiss and Ankeny, 2016, section 5). The general aim of this community is to promote quantitative, statistical evidence for medical decision-making at the clinical level over qualitative evidence such as case reports and expert consensus, which are considered fraught with bias and uncontrolled confounding factors. By contrast, these potentials for misleading evidence can be (better) controlled in RCTs and the meta-analysis thereof.

Although there is increasing philosophical analysis and critique of various aspects of the EBM framework—see, e.g., Worrall (2007) for a review—meta-analysis has been identified as one such aspect in need of greater attention (Mebius et al, 2016). As one of the first philosophical analysis to focus on meta-analysis in particular, Stegenga (2011) raises at least three sorts of important issues. First, meta-analysts seem to have many arbitrary choices to make in order to complete their work, raising the specter of impotence or conventionalism if their conclusions depend on the details of these choices. Second, meta-analyses typically focus only on statistical data and so neglect other important types of evidence, e.g., mechanistic evidence, which may be especially relevant for policy and decision-making. Third, meta-analyses are not immune to the damaging effects of publication bias, p-hacking, experimenter degrees of freedom, and other questionable research practices whose trace in the published literature can be difficult to detect, thereby impugning its objectivity and claim (within EBM) to evidential superiority.

In this special issue, Holman (2018) explicitly defends meta-analysis, pointing out that once one conceives of it as a developing process rather than a static

technique, some of the charges against it are no longer apropos, while others can be resolved after periods of problem solving. Wüthrich and Steele (2018) are also interested in defending the utility of meta-analytic methods, especially in cases where automation becomes necessary due to the super-large-scale data involved. In these cases, careful problem solving cannot be done on a case-by-case basis and must instead be built into the automated aggregation algorithm; they suggest amenability to robustness analysis (section 2.3) is an important aspect of design.

2.3 Robustness/Sensitivity Analysis

The notion of robustness comes in different shapes and forms (Lloyd, 2015; Jones, 2018; Schupbach, 2018; Staley, 2004; Weisberg, 2006; Woodward, 2006). Weisberg delineates a four-step procedure of mathematical robustness analysis. In the first step, a group of models is examined to determine if they all make the same prediction and if they are sufficiently diverse. The second step aims at finding the common structure which generates the prediction. The third step is the relating of the model predictions to the real world (on which cf. section 2.4). The final step is to conduct an analysis to investigate conditions which defeat the prediction.

Wilde and Parkkinen (2018) expand on Weisberg's first step. But rather than constraining themselves to a robustness of mathematical models, they also consider varying modeling assumptions, detection methods and experimental set-ups (such as experimental species)—see Culp (1995) for more on robustness and experimental set-ups. Woodward (2006) calls this derivational robustness, whose confirmational value has recently been argued for by Eronen (2015); Kuorikoski et al (2012); Lehtinen (2018). Again we find a notion of diversity playing a key role: the more divergent the group of experimental species for which a robust result is found, the less likely the observations are owed to idiosyncrasies of the individual species—in other words, the more likely it is that the observed phenomenon is also manifest in the species of interest. It is an intuition concerning the weight of the *variety of evidence* (section 2.1) which confirms the hypothesis of interest via extrapolation (section 2.4). Schupbach (2018) also traces intuitions regarding extrapolation back to intuitions concerning variety of evidence expounded in Horwich (1982).

Derivational robustness is also the most basic form of the fourth step in Weisberg's procedure. Intuitively, changing a single or few parameter values by a fraction only changes the conclusions minimally, if at all. (Raerinne (2013) is careful to distinguish procedures like sensitivity analysis from other types of robustness analysis.) In mathematics, this notion is often formalized as the continuity of a function. By contrast, in situations with manifold discontinuities, reliable predictions become all but impossible as chaos reigns. In between these extreme scenarios, chaos theory may be applied in situations with *some* discontinuities.

In scientific practice, such robustness analysis is often carried out via the execution of Monte Carlo algorithms on a computer (Lagoa and Barmish, 2002; Rubinstein and Kroese, 2016). Roughly speaking, such algorithms probabilistically explore how one's conclusions change for different input and parameter values. Wüthrich and Steele (2018) argue that evidence amalgamation algorithms ought to be assessed by considering the kind of robustness analysis, thusly understood, that can be performed; the possibility space associated with the robustness analysis is revealing of the basic structure of the algorithm.

By contrast, Kao (2018) employs a notion of robustness to support the development or discovery of a theory or hypothesis rather than its evaluation. She points out that attempting to generalize specific hypotheses for concrete domains to further areas of application is a viable heuristic research strategy which sheds light on the possible unifying work that a generalized hypothesis may do. Furthermore, determining boundaries limiting the scope of such generalized hypotheses may help our understanding of the epistemic values of scientific hypotheses. The search for boundaries outside which the theory no longer holds is a search for defeaters—the fourth step in Weisberg's robustness analysis.

2.4 Extrapolation and External Validity

There are many examples in the sciences in which evidence is collected to support a hypothesis about a certain system, but this system is not directly accessible for financial, ethical, or technical reasons. In these cases, surrogate model systems stand in as test objects in experiments, and experimental findings are then transferred from the model system to the target system. This type of inference is called *extrapolation*.

Obviously, the concept of extrapolation brings with it a host of interesting and difficult questions, such as "what is a model?" and "how do models represent?", touching philosophical issues such as the problem of induction (as some gap between model and target must be leaped over), the status of universal laws (as the model and target may not fall under the same law), and the nature of causation (since it is almost always causal knowledge that is extrapolated). Another persistent problem is the question of what degree of contextualization is needed for a successful transfer of knowledge, especially in the biomedical sciences with large, complex, and dynamical systems that come with all kinds of redundant mechanisms. Moreover, the philosophical debates around the nature of similarity, relevance, and analogy cannot be neglected here, either.

The concept of extrapolation is tightly entangled with questions about internal and external validity. When an experiment is performed on an animal model or a study conducted on a test population (which in medical settings is sometimes quite small), a causal claim can only be established if the experimental set-up or the study design is judged internally valid, i.e., it was really

C which caused E in the setting M. (See also the debate around randomized controlled trials described in sections 2.2 and 2.6.) When the experiment or study goes beyond M and is deemed externally valid in that the causal link between C and E does not only hold in M alone, then causal knowledge about M might be justifiably extrapolated to some distinct target setting/population T. The question of whether this transfer is permissible for a given M-T pair is dubbed *the problem of extrapolation*—e.g., transferring causal knowledge about a drug's effects from animal models to humans might be sensitive to certain particularities of the study setting (such as age, co-morbidity, etc. in the study's sample).

In his discussion of mechanistic reasoning for the purpose of extrapolation, Steel (2008, p. 78) presents the following additional challenge any viable account of extrapolation ought to address:

> [A]dditional information about the similarity between the model and the target—for instance, that the relevant mechanisms are the same in both—is needed to justify the extrapolation. The extrapolator's circle is the challenge of explaining how we could acquire this additional information, given the limitations on what we can know about the target. In other words, it needs to be explained how we could know that the model and the target are similar in causally relevant respects without already knowing the causal relationship in the target.

(See also Guala (2010).) Different proposals have been put forward to evade this circle, including mechanistic reasoning (e.g., Steel (2008) on comparative process tracing) or analogical reasoning in a Bayesian framework (Poellinger, 2018). Reiss's contribution in this issue (2018) offers an overview of strategies to tackle the problem of extrapolation, while Reiss himself proposes an alternative pragmatist, contextualist perspective on evidential support for an inaccessible target system. Comparative process tracing is picked up again by Wilde and Parkkinen (2018) as a way of basing extrapolation on mechanistic reasoning to show how both probabilistic and mechanistic information can be transferred at once when establishing a causal claim about a human target population. Frank (2018) adds an ethical dimension to the discussion by shifting the focus to reasoning under uncertainty when results from locally validated models of climate change-related economic effects are extrapolated.

2.5 Evidence Amalgamation and Theory Development

It is natural, as many of the previous sections have done, to consider evidence amalgamation as pertaining to the confirmation of scientific theories and hypotheses. But beyond its role in the logic of scientific justification, it also figures in the logic of scientific discovery. (Here, "logic" is understood in a broad sense as a form of rational inquiry.) Whewell (1840) was one of the first to distinguish the process of conceiving of a theory or hypothesis from the

bulk of the establishment of its empirical support. He identified three stages to scientific inquiry, as we would call it today:[1]

1. the "happy thought," or the event of the novel insight or idea properly so called;
2. the "colligation" of facts and ideas, or the further formation and maturation of the happy thought by its relation and integration into other ideas and known data; and
3. the verification of the colligation, meaning the judgment of its explanatory and predictive power and its simplicity compared with its "consilience," that is, its unifying range of application.

The first, and possibly also the second, of these steps are included in the modern conception of scientific discovery.[2] Clearly, the role of amalgamating diverse evidence figures most centrally in the second and third steps, hence in discovery itself insofar as the second step is included therein. In this special issue, Kao (2018) focuses on this second step with the example of the development of the early quantum theory, showing how versions of the quantum hypothesis guided further experimental results, which in turn constrained their scope.

For much of the twentieth century, however, most (but not all) philosophers saw the process of scientific discovery as an essentially creative phenomenon beyond the purview of philosophy of science, whose task was to delineate the normative constraints on the scientific endeavor. Only in the 1970s did philosophers devote noticeably more attention to it (Schickore, 2014). An early exception was Hanson (1958, 1960, 1965), who articulated a theory of discovery as abduction, whereby diverse phenomena, particularly those unexplained, are unified as following from the truth of a certain hypothesis. To the extent that this hypothesis amalgamates diverse evidence, it is a good candidate for further investigation. Although Hanson cited Charles Sanders Peirce as inspiration, there is clear continuity with Whewell's conception. Further, Magnani (2001, 2009) has later emphasized that abduction needn't be used to select a single hypothesis once and for all—that is, as a mode of inference properly so called—but can also be used creatively to generate or refine further hypotheses.[3] Whether one views this abduction as a "logic" in any relevant sense, such unifying or abductive reasoning concerns the developments and pursuits of hypotheses and theories rather than their direct support or justification. In this special issue, Bertolaso and Sterpetti (2018) follow this line of argument concerning how cancer researchers should pursue theories of carcinogenesis according to their plausibility.

[1] Whewell referred to the whole process as that of scientific discovery, reflecting the older usage of that sense of "discovery" as broad inquiry.

[2] Cf. Laudan (1980), who prefers to distinguish all three, calling the second the context of pursuit.

[3] See also Schaffner (1993), who sees this use of abduction as a "weak" evaluation procedure, providing more evidence of scientific promise rather than confirmation.

2.6 Causal Inference

After—and despite—Russell's famous, skeptical wholesale rejection of the concept of causation (Russell, 1912), the past hundred years have seen a surge in approaches towards (more or less) formally explicating cause-effect relationships. Reductive or non-reductive in nature, none of the explications or definitions rests on a single marker: probabilistic accounts (Suppes, 1970; Reichenbach, 1971) combine information about two events' correlational and temporal relations, prominent causal graph accounts (Pearl, 2000; Spirtes et al, 2000; Woodward, 2003) aim at the integration of knowledge about probabilistic relations and underlying (spatio-temporal) mechanisms, and process theorists about causation build on descriptions of physical mechanisms and difference-making information, either in terms of energy transfer (Salmon, 1984) or a system's counterfactual development (Dowe, 2009). In applied settings, several of these can be combined: the Russo-Williamson Thesis (Russo and Williamson, 2007) captures the desideratum to enrich probabilistic data (as gleaned from RCTs) with mechanistic information towards the establishment of justifiable causal claims in medicine—for a discussion of the thesis see Wilde and Parkkinen (2018) in this issue.

Let us briefly take a closer look at the causal graph account. The causal modeler might start by synthesizing probabilistic data from databases, background information, common sense, and expert knowledge. These different sources might provide both structural/mechanistic as well as parametric (or also distributional) information about the relations between different factors. Causal learning algorithms can help in discovering unsuspected relationships, but the integration of expert knowledge becomes even more important as the number of investigated variables grows and computational tractability quickly gets out of hand. Moreover, a larger causal theory might be constructed as a patchwork theory by combining smaller, local structures collected from different experiments, different research groups, or even different branches of science (Mayo-Wilson, 2018). When dynamic information is added to the mix in the aggregation of time-series data, a host of new inferential problems arises (as discussed in Danks and Plis 2018). Interesting conceptual and computational innovation towards solutions to these problems and in synthesizing ideas is increasingly driven by exchange between computer scientists and philosophers.

But, if one takes a skeptical stance towards such a monistic explication of cause and effect, advocating instead for a pluralist explication, one should be prepared to amalgamate causal evidence in a fundamentally particularist way. For example, in her critique of the causal graph approach, Cartwright (2004, pp. 814–815) dismisses this "monolithic" account as too formal and "thin" and advocates richer terminology closer to experimental practice:

> [T]here is an untold variety of quantities that can be involved in laws, so too there is an untold variety of causal relations. Nature is rife with very specific causal laws involving these causal relations, laws that we represent most immediately using content-rich causal verbs: the pistons *compress* the air in the carburetor chamber, the sun *attracts* the planets,

> the loss of skill among long-term unemployed workers *discourages* firms from opening new jobs. . . . These are genuine facts, but more concrete than those reported in claims that use only the abstract vocabulary of "cause" and "prevent." If we overlook this, we will lose a vast amount of information that we otherwise possess: important, useful information that can help us with crucial questions of design and control.

Thus, for the causal pluralist, amalgamation of varied evidence is necessary to preserve rich information about the causal system under investigation. (See also Reiss's sketch of his pragmatist-contextualist theory of evidence for causal inference in Reiss 2018.)

3 Contributions

3.1 Claveau and Grenier on the Surprising Failure of the Variety of Evidence Thesis

Claveau and Grenier (2018) follows up on Claveau (2013) in offering more nuanced understandings of varied evidence in the sense of Bovens and Hartmann (2003) and section 2.1. In this tradition, instruments are assumed to be either α) fully reliable, delivering perfect information about what they measure, upon which one can deduce infallibly the consequences, or β) fully unreliable in that they do not provide any information whatsoever (reports from these instruments are random and mutually independent). As Claveau (2013) notes, real scientific instruments do not work like this.

Claveau (2013) models unreliable instruments as biased instruments that either always create reports that the hypothesis of interest is true or always create reports that the hypothesis of interest is false. Eliminating the testable consequence variables from the models, Claveau (2013, Section 4) shows that a natural formalization of the Variety of Evidence Thesis holds in his models.

In a second step, he drops the assumption of Bovens and Hartmann that the reliability variables are either fully independent or not independent at all. He discovers situations in which reliability variables have intermediate degrees of independence, for which the Variety of Evidence Thesis fails.

In their contribution, Claveau and Grenier (2018) extend this analysis by first re-introducing the testable-consequence variables into the models and then dropping the assumptions that these variables are either fully dependent or fully independent. These, so far, most general models, are depicted in Figure 2.

Within this general model, they show that a variety of possible understandings of the Variety of Evidence Thesis fail for a large set of plausible priors. In a natural sense, the set in which the Variety of Evidence Thesis fails is much larger in Claveau and Grenier (2018) than it is in Claveau (2013). Consequently, the conclusions drawn for the thesis are less favorable in Claveau and Grenier (2018) than in Claveau (2013).

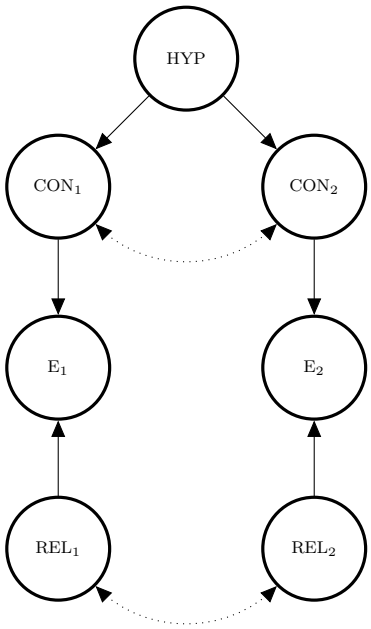

Fig. 2 The most general model of Claveau and Grenier (2018) with evidence variables and consequence variables that are dependent to some degree.

3.2 Heesen et al in Praise of Methodological Triangulation

Heesen et al (2018) are interested in the work of Du Bois (1996/1899) concerning the question, "What do Negroes earn?" (To avoid anachronism, they follow Du Bois' terminology in using "Negro" rather than the more contemporary "African-American.") To make progress in answering this question for households, they consider four methods for gathering information: 1) conducting interviews, 2) combining the average income for the professions represented in a given household, 3) using family members' estimations of time lost to work, given their occupation, and 4) judging the appearance of the home and occupants, rent paid, and the presence of lodgers. They are then interested in how to aggregate answers obtained from these different methods.

Heesen et al (2018) delineate i) a purist strategy—single out a method and believe whatever this method finds—from ii) a triangulation strategy—believe what most methods find and break ties by randomly picking from best supported findings. They show that the triangulation strategy has a greater probability of yielding the correct answer than the purist strategy for $m \geq 3$ methods and $n \geq 2$ possible answers. Furthermore, this probability for the triangulation strategy increases with the number of methods m. Their model hence underwrites an understanding of the Variety of Evidence Thesis in which variety is understood as the employment of a variety of methods to support one's inferences.

Interestingly, their model can almost be understood in terms of the Bovens and Hartmann framework. The first ingredient of such a model is an n-ary propositional variable HYP. Second, one considers m different propositional consequence variables CON which can take the same values as HYP. The relevant conditional probabilities satisfy $P(Con = con|HYP = hyp) = 1$ if and only if $con = hyp$—that is, the consequences are perfect indicators of the hypothesis. Every consequence variable CON has a different single child E of arity n for which $P(E = e_i|Con = con_i) > P(E = e_j|Con = con_i)$ holds for all $j \neq i$, see the left-hand graph in Figure 3. Landes (2018, Section 6) shows that, under suitable ceteris paribus conditions, the posterior Bayesian probability in the correct answer increases in m when $n \geq 2$, even if the consequences are not perfect indicators of the hypothesis.

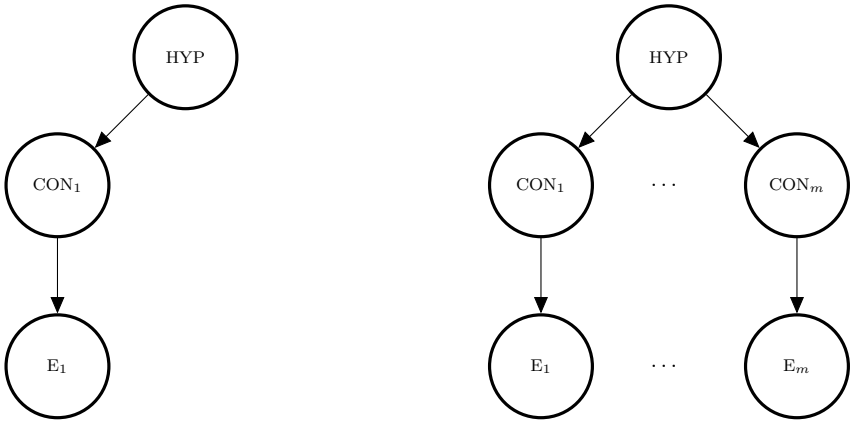

Fig. 3 The triangulation and purist strategies described by Heesen et al (2018) within the Bovens and Hartmann (2003) framework.

The crucial difference between these two approaches is how varied evidence is used to update beliefs. Heesen et al (2018) employ a triangulation strategy while Bovens and Hartmann (2003) update using Bayesian conditionalization. Those with very strong intuitions for the Variety of Evidence Thesis may thus feel compelled to give up on Bayesian updating.

Since the proofs of Heesen et al (2018) heavily draw on voting theory (List and Goodin, 2001), we have hence connected voting theory to variety of evidence reasoning. It is tempting to speculate about deep connections between the two. These connections, if they exist, are as yet unexplored.

3.3 Holman in Defense of Meta-analysis

In his contribution, Holman (2018) mounts a spirited defense of meta-analysis as a flexible toolkit to amalgamate data for (medical) decision making, against

criticisms from Stegenga (2011), Jukola (2015), Romero (2016), Holman and Bruner (2017), and Jukola (2017). He argues that worries raised for meta-analysis can be dealt with by the rich meta-analytic toolkit and that worries concerning the objectivity of meta-analysis either pose a major threat to all other forms of evidence amalgamation, too, or impose an unreasonably high standard. Crucially, though, he understands his argument as a defense of, but not an argument for, meta-analysis.

In more detail, Holman (2018) understands meta-analysis as an ongoing process, much like other scientific methods. The process is ongoing in two respects: i) the available evidence accumulates over time, which may bring resolution to disputed issues, and ii) the meta-analytic tools themselves as well as tools to assess meta-analyses improve over time. While there may be disagreement between competing meta-analyses at any given time, there is good hope that the disagreement will disappear over time either due to new evidence coming to light or by detecting virtues and flaws in the present meta-analyses. So, meta-analysis is not a tool which meets the unreasonable standard of instant elimination of disagreement but it is rather the ongoing process of scientific inquiry which leads to the eventual elimination of disagreement over time.

Next, he turns to issues arising from considering meta-analysis as a social practice, acknowledging the importance of assessing the impact of industry funding and publication bias. He first considers concerns raised by Jukola (2015, 2017), regarding worryingly skewed data from the FDA. She hence pointed out that even a well-conducted meta-analysis on the skewed data produces a systematically biased result. Holman (2018) addresses this worry by pointing to meta-analytic tools for detecting skewed data—PRISMA and funnel plots. Recent computer simulations by Romero (2016) make a point similar to Jukola's. Holman (2018) argues that applying the p-uniform technique in Romero's implicit meta-analysis would have been the best practice to mitigate effects of biased publications.

Holman (2018) thinks that it is rational in practical terms to violate the Principle of Total Evidence (Carnap, 1947) in medical inference. Firstly, there are many cases in which ignoring non-RCT studies has produced the right results, by the light of history. Secondly, research in cognitive psychology suggest that we tend to interpret complex information in ways that fit our view of the world. So, focusing on smaller bodies of high quality evidence reduces the risk of biased interpretations.

3.4 Vieland and Chang on Conceptual Problems for Classical Meta-Analysis

The Fisherian approach to classical statistics determines what one might call the "evidential bearing" of a data set *against* a hypothesis according to how extreme the data set is compared with what is expected under the hypothesis. In other words, the data are interpreted as evidence against an hypothesis to the extent that the data do not fit it. For example, in an experiment of ten independent coin flips, there is evidence against the hypothesis that the coin

is fair to the extent that the number of heads (say) differs from five. This evidence against the hypothesis of fairness is often quantified by a *p-value*, the probability of measuring data at least as extreme as the data actually observed, if the hypothesis were true.[4] The question then arises: how does one amalgamate the evidence across many statistical studies against a hypothesis? If the p-value measures the evidence, then this question becomes: how should one combine p-values? Doing so justifiably is the goal of classical meta-analysis.

In their contribution, Vieland and Chang (2018) draw philosophers' attention to neglected conceptual problems involved in these procedures for classical meta-analysis, even when one focuses on the simplest case of multiple independent replicates of the same study type, design, and size. They show that three different procedures for meta-analysis—what they call p-averaging, r-averaging, and replication counting (along with their likelihood-ratio-measure analogs)—all have problems being justified as an evidence amalgamation procedure or conflict with what they call the "measurement amalgamation principle" (MAP), that the inputs and outputs of an amalgamation procedure should be on the same measurement scale, with a meaningful interpretation vis-à-vis the evidence.

P-averaging, which they attribute to Fisher (1925), satisfies the first clause of MAP, as it outputs the p-value associated with a certain statistic proportional to the sum of the logarithms of the p-values of the input studies. But there are reasons to doubt that it satisfies the second clause, calling into question whether the p-value was a measure of evidence in a single study in the first place. In r-averaging, on the other hand, certain aspects of the data themselves are first amalgamated and then the p-value is calculated for this larger data set, resulting in a procedure that fails to satisfy the first clause of MAP. Finally, they consider binary classification schemes, which coarse-grain the evidence provided by statistical studies into those that support the rejection of a certain hypothesis or not, depending on whether their p-values are below or above a fixed threshold, respectively. Simply put, these schemes are so coarse that it is hard to justify how they could be reliable procedures for amalgamating evidence.

While Vieland and Chang (2018) do not offer a positive proposal to overcome these problems, they describe in some detail various constraints on potential solutions and directions for future research, such as the connection with parameter estimation.

3.5 Wüthrich and Steele on Automated Evidence Aggregation

Wüthrich and Steele (2018) are concerned with evaluating procedures for aggregating ever-larger bodies of evidence. The rapid growth of available data

[4] Technically, the p-value is of a statistic of the data, which orders the possible data by increasing extremeness. Typically data is taken to be extreme to the extent that it has low probability (or low probability density).

will eventually require the application of automated evidence aggregators implemented as computer algorithms. Recent work by Hunter and Williams (2012) in automated aggregation of medical data serves as a case study to their philosophical questions.

They are particularly interested in how to choose the proper *extent* of automation. They first argue that it is non-trivial to determine criteria which allow to specify an/the appropriate extent of automation in any particular application. However, they do put forward and defend one criterion for assessing the reliability of evidence aggregation algorithms: the capacity to perform adequate *robustness analysis*.

But why is such a criterion necessary? Can we not simply assess the track record of these algorithms? Unfortunately, this is often not possible in the medical domain when one attempts to determine a priori the best treatment option for a particular patient at a given time. There are at least three reasons why this may not be possible: a) there is no way to go back in time and check whether the other treatment options would have been any better, b) while the data were correctly amalgamated, it was the raw data themselves which were misleading, and c) the available good data were correctly assessed but the patient belongs to a rare sub-group in which the indicated treatment is—surprisingly—a lot less beneficial than in the general population.

So, by "robustness analysis" Wüthrich and Steele (2018) refer to a relatively simple notion of robustness in terms of varying parameter values or "dial settings." They argue that designing large-scale evidence aggregation algorithms with the possibility of such robustness analysis has two crucial advantages: it becomes clear from early on in the design process i) which (types of) uncertainty can be subjected to robustness analysis by "turning the dial" and ii) which types of uncertainty cannot be subjected to robustness analysis because there is no "dial" one could turn. This directs attention to the structure of the algorithm—the choice and organization of the dials. Thus, it also increases transparency of the inner workings of the algorithm.

Wüthrich and Steele (2018) point out that in actual practice, further complications arise due to limited (computational and other data processing) resources. An evidence aggregation algorithm may produce better results by a deeper analysis of a smaller data set than by a shallower analysis of a larger data set. These complications notwithstanding, they argue that the assessment of an automated evidence aggregator is to be assessed by focusing on robustness analyses.

3.6 Wilde and Parkkinen on Extrapolation and the Russo-Williamson Thesis

In their contribution, Wilde and Parkkinen (2018) take a closer look at the so-called Russo-Williamson Thesis (RWT), which can be understood as a programmatic claim about the importance of evidential variety in causal assessment. In its strong formulation, the thesis says that, for the establishment of a causal claim, both knowledge of the existence of a causal mechanism and

knowledge of a suitable correlation must be established (Clarke et al, 2014, p. 343). This thesis, so its authors claim, is supported by research practice, and especially by work done at the *International Agency for Research on Cancer* (IARC), where multiple research groups work on different sources of evidence for or against the carcinogenicity of various substances. When separate research groups arrive at a conclusion, these findings are combined in order to categorize the investigated substance as carcinogenic to humans (category 1), probably carcinogenic to humans (category 2A), possibly carcinogenic to humans (category 2B), not classifiable as to its carcinogenicity to humans (category 3), or probably not carcinogenic to humans (category 4). Although the IARC's way of reasoning can be understood to be in accordance with the RWT, various seeming counterexamples to this practice have been pointed out in the recent discussion. Wilde and Parkkinen briefly mention the case of the carcinogenicity of processed meat, where the classification was seemingly based on correlational evidence alone. One answer to this deviation in practice might be a call for correction (i.e., emphasizing the importance of mechanistic evidence, as in Leuridan and Weber (2011)), while an alternative answer might be a reinterpretation of the the case: Clarke et al (2014, p. 343) point out that a study of sufficient quality may provide correlational information and at the same time rule out the possibility of confounding and bias.

Wilde and Parkkinen (2018) investigate a second case, the causal link between benzo[a]pyrene and cancer in humans. In this case, benzo[a]pyrene was classified by the IARC as carcinogenic to humans without an established correlation in humans. In defense of IARC's practice and the RWT, they show how evidence from animal studies can possibly yield both correlational and mechanistic information about humans through mechanism-based extrapolation from experimental animal models to the human target population (Steel, 2008, p. 85). In this special case, extrapolation is supported by particularly robust evidence from animal studies. The IARC protocols report eight species of dissimilar non-human model animals, with the experimental results remaining stable across species and therefore likely to be due to a common causal feature of all experimental set-ups: the underlying phenomenon, i.e., the carcinogenicity of benzo[a]pyrene. In this sense, the IARC practice can be understood as being in accordance with the RWT again, after all: evidence from animal models can be reliably transferred from animals to humans via mechanism-based extrapolation and thus provide both probabilistic and mechanistic information for the establishment of a causal claim about the human target population.

3.7 Frank on the Ethics of Economic Extrapolation from Uncertain Climate Models

Frank (2018) discusses ethical issues involving risk reporting where scientists extrapolate from locally validated models. In particular, he is interested in extrapolated estimates of cost and damage within so-called Integrated Assessment Models of the economic effects of climate change. The varied evidence

that goes into constructing these models—e.g., the functional dependence of warming on atmospheric carbon dioxide, the ability and efficacy of technology to allow economies to adapt to and mitigate climate warming, and the effects of more extreme warming on global security and scarcity of resources—entail high uncertainty about these models' predictions, especially since that evidence is confined to relatively low amounts of global warming. Frank (2018) expounds on the reasoning of Weitzman (2009), who critically observes that such extrapolations have so far been based on analytic tractability rather than some more evidentially justified reason. If the aforementioned high uncertainty is taken in account, it tends to lead to "fat tails" in the probability distributions for catastrophic events, such as the end of civilization as we know it.[5] This can then lead to negatively infinite, or arbitrarily large, expected value for climate mitigation risks. The disconcerting conclusion that a rational actor should do anything and everything to mitigate such risks could perhaps be avoided by placing an upper bound thereon using the technique of a *statistical value of a life* multiplied by the world population, but this is controversial because its coarseness has yet to be furnished with a justification.

Nevertheless, Frank (2018) argues that the Weitzman (2009) approach should be taken seriously. Two ethical norms give it support: first, that scientists should be sensitive to both the epistemic and non-epistemic consequences of their research conclusions, managing risk therefrom accordingly, and second, that their own uncertainty about their conclusions should be made especially transparent for policy makers. Insofar as the fat-tailed approach is epistemically permissible, researchers should represent that extreme uncertainty to policy makers, on precautionary grounds (Steel, 2014). The risks associated with overly aggressive climate change mitigation are simply dwarfed by those with the collapse of civilization. While there are still uncertainties that remain about how to balance epistemic and ethical values and risks, the argument for the fat-tailed approach remains plausible.

3.8 Reiss Against External Validity

Reiss (2018) argues that understanding evidential reasoning in terms of external validity may lead to poor inferential practice by encouraging the search for epistemically easily accessible models, which are then used as the basis for extrapolations. The alternative he proposes involves the attempt to understand evidential reasoning about targets more directly. In criticizing evidential reasoning in terms of external validity, Reiss first introduces the problem of extrapolation, i.e., the question how knowledge of a causal relationship (about a model system) gained in experimental circumstances can justifiably be transferred (extrapolated) to less accessible circumstances (the target system, often the target population). He then presents solution strategies discussed in the literature to show that evidential reasoning based on external validity encourages

[5] More technically, probability distributions with "fat tails" are ones over the real line that asymptotically decay sub-exponentially, leading to undefined moments.

foundationalist thinking about scientific inference: Learning whether C and E are causally related in target system T requires taking the detour through the experimentally accessible model system M. Yet, according to Reiss, foundationalism only offers unsatisfactory answers to questions about a) *how the basic beliefs (inferences from experimental data) are justified*, and b) *how justified basic beliefs about M also lend justification to beliefs about T*.

As an alternative to external validity-based scientific reasoning, Reiss proposes a pragmatist-contextualist perspective (Reiss, 2015): reasoning about a target system T should center around the hypothesis and begin with the question of what evidence is required in order to establish this very hypothesis. In Reiss's pragmatist framework, the hypotheses together with their context—domain-specific information, purpose of the inquiry, norms, etc.—determine what kind of evidence is needed to support the target causal claim, and what counts as justification in this situation. Contextualism in this sense facilitates reasoning from models without external validity. In a series of examples taken from cancer research and IARC practice, Reiss illustrates the fruitfulness of the contextualist approach: Experiments might suggest hypotheses in the discovery stage, animal experiments may provide direct support for a hypothesis once a suitable (domain-specific) *bridge principle* is available, animal experiments might also refine hypotheses if they suggest a more specific causal relationship, and analogies can be exploited if knowledge about the mechanisms involved is available. All these inferential patterns facilitate the integration of evidence in support of a causal hypothesis but do so without extrapolation or judgments of external validity.

3.9 Bertolaso and Sterpetti on Plausibility and Cancer Research

Bertolaso and Sterpetti (2018) argue that the analytic view of theory development (Cellucci, 2016, 2017) can shed novel light on how varied evidence influences this development. According to this view, one ranks hypotheses by their *plausibility* according to the following procedure: deduce conclusions from the hypothesis, then compare the range of these conclusions with other provisionally accepted hypotheses and data. Those that yield contradictions are less plausible and so are provisionally rejected, while those that remain are provisionally accepted. They argue that the plausibility concept is distinct from probability, because it arises from the balance of reasoned arguments for an hypothesis from diverse evidence, which is a non-quantitative relation. This better explains the actual pronouncements of scientists regarding their theories, who rarely give probabilistic estimates for them, and may go some way to explain how different researchers assign different prior probabilities to hypotheses. Because plausibility depends on an argumentative structure, they emphasize that development cannot be fully automated without the input of researchers, contra big data advocates such as Gagneur et al (2017). This in turn has implications for the debate between frequentists and Bayesians about RCTs.

Bertolaso and Sterpetti (2018) illustrate their view using the current state of the art in theories of carcinogenesis. The dominant Somatic Mutation Theory (SMT) posits that the cause of cancer is accumulated mutations in a cell that lead to proliferation instead of the default state of quiescence. The minority Tissue Organization Field Theory (TOFT), on the other hand, posits that the cause of cancer is abnormally imbalanced interactions among adjacent cells and tissues, whose default state is rather proliferation. SMT and TOFT are rivals insofar as they posit different default states for cells, resulting in different assessments of the mutations found in cancer cells: either a cause (SMT) or an effect (TOFT). But there is difficulty finding decisive evidence for or against either theory. In the case of SMT, for example, researchers have used big data sets to search for *driver* mutations, the mutations allegedly responsible for carcinogenesis. However, the interpretation of the data as such evidence requires accepting the SMT in the first place. So this search is not best understood as for a probabilistic confirmation of SMT over TOFT, but as the articulation of a plausibility argument for it; conversely, proponents of TOFT mount arguments according to which the data can be explained by their theory.

3.10 Kao on Unification for Theory Development

Unification, as the process of bringing together a disparate set of phenomena under a common understanding, has long been recognized as one strategy for explanation, as has the abductive inference to the hypothesis or theory providing that understanding as a method of confirmation or justification. Kao (2018) reminds us that seeking such unification also provides a strategy for theory development—cf. section 2.5. Here Kao appeals to the second stage of the inductive method of Whewell (1840, 1858), the "colligation of facts" by which an idea is elaborated and delineated. The initial articulation of an idea does not always yield a precise hypothesis or entail a definite theory. Attempts to unify phenomena under an idea therefore help delineate the scope and content of an hypothesis or theory: how broadly applicable the idea is, and what it is in detail. As a heuristic, unification is distinct from confirmation or a commitment to any broad unity of science, because *failures* to unify can delineate the scope and content of a theory, too. Finally, Kao compares this sort of unification with uses of robustness analysis to develop a theory rather than confirm it.

To illustrate this thesis, Kao examines the quantum hypothesis in the early quantum theory of Planck (1967/1900), Einstein (1967/1905, 1907), and Bohr (1913) from 1900–1913. As Planck (1967/1900) originally formulated it in the context of describing the blackbody radiation spectrum, the quantum hypothesis was ambiguous, as an assumption of discreteness, between the states of elemental constituents of the blackbody and their energy. Einstein (1967/1905) then extended this idea to electromagnetic radiation itself, rather than just the material part of a matter-radiation system as Planck (1967/1900) had applied

it. In the other direction—that is, regarding matter only without radiation—Einstein (1907) applied it to the energy spectrum of atoms in diamond. Now, the quantum hypothesis entails the existence of a constant, h, in units of which the energy of vibratory phenomena (at some frequency ν) are discretized, but the scope of the universality of this constant was unclear. It was Bohr (1913) who extended this scope to provide the information needed to calculate the characteristic size of atoms. However, the proposed extension of Debye and Sommerfeld (1913) to quantize the duration of an energy exchange process in an electric field was not born out by experiments. This showed the quantum hypothesis was delimited to the descriptions of systems rather than temporal processes.

3.11 Danks and Plis on Amalgamating Evidence of Dynamics

Danks and Plis (2018) turn their attention to the problem of amalgamating statistical time-series data. The authors note that in many cases where large datasets are compiled over long durations by different teams possibly measuring different parameters with different methods at different time scales, merging the resulting databases is highly desirable but hampered by technical problems. One step towards amalgamating information about the behavior of complex, dynamical systems (as, e.g., in neuroscience, econometrics, climatology, etc.) is to integrate different studies not at the level of data but, in a move to sidestep some evidence amalgamation challenges, at the structural level precisely because the underlying causal structure remains invariant across studies. However, how well structural causal knowledge can be amalgamated in the end relies on how well the causal structure can be extracted from the raw data. In the case of extracting such structural information from time-series data, two challenges arise on which the authors focus: 1) there might be a mismatch between the measurement timescale and the causal timescale, and 2) latent, unobserved variables might confound dynamical systems over temporal durations.

Danks and Plis illustrate the first challenge with an example from brain research, where measurement timescales of different methods deviate significantly: magnetoencephalography takes a measurement each 1 ms, while fMRI data is sampled with one measurement every 2000 ms. As a result, causal structures with different underlying causal time-steps will be extracted from their respective databases. Danks and Plis visualize the loss of information (i.e., "disconnectedness") through undersampling in causal graphs and discuss algorithmic approaches towards learning causal timescale structure from measurement timescale data (for known and unknown undersampling rates).

The second challenge consists in amalgamating different causal structures into one global structure in the presence of unobserved latent influences. In particular, if the sets of measured variables differ between studies, we know that not all variables are measured in all experiments. Most importantly, such unmeasured variables might act as confounders. Yet, recovering latent struc-

tures from time-series data is made even more difficult by the problem of underdetermination: How many latent variables? How many self-loops? How many time-steps from cause to effect? To tackle this second challenge, Danks and Plis propose to extend the causal structures extracted from time-series data to minimally enriched ("simplest") networks expressing causal influence through latent variables over temporal distances. Danks and Plis discuss concrete algorithmic solutions that can finally help in merging the resulting causal structures.

3.12 Mayo-Wilson on Causal Collages and Piecemeal Experimentation

Mayo-Wilson (2018) also focuses on causal learning, in particular on how to combine the results of smaller studies into larger causal collages. Such practice is quite common in medicine and in the social sciences where researchers often confine their attention to a limited set of variables first and combine the inferred causal knowledge with other researchers' results later. Mayo-Wilson's discussion builds on earlier work (Mayo-Wilson, 2011, 2014) showing that in combining causal knowledge inferred from observational studies (which might in many instances be the only option due to ethical, financial, or technical constraints), information might be lost—in other words, causal theories might be significantly underdetermined by evidence when observational data is gathered *piecemeal* in contrast to comeasuring more (or all) variables at once. Mayo-Wilson (2018) argues that this "problem of piecemeal induction" persists even when *interventions* in a given setting are possible. He then investigates three interrelated questions regarding this problem.

Firstly, Mayo-Wilson asks what type of information (and how much) is lost in the piecemeal aggregation of evidence for the purpose of constructing a causal theory. As soon as experimentation is possible, no information is lost regarding the direction of the "causal flow." Nevertheless, depending on the connectedness of the true causal structure, quite some information regarding the presence (or absence) of causal connections can be lost.

Yet, in which cases does no information loss occur when merging the results of multiple studies? Mayo-Wilson (2018) shows that underdetermination of the inferred causal theory can be eliminated if the graph of the true causal structure contains relatively few edges.

Lastly, Mayo-Wilson asks how often the problem of underdetermination arises. Unfortunately, experimental interventions do not reduce the underdetermination rate (in contrast to inference from observational data) when the number of variables in the causal graph becomes large (as is the case in many real-world settings, especially in medicine or sociology).

Balancing his skeptical outlook on the fruitfulness of interventions for causal learning, Mayo-Wilson remarks that scientists usually build their investigations on more than just probabilistic knowledge or facts about conditional independencies: Most importantly, domain-specific knowledge (such as plausibility constraints) might be available and useful in reducing underdetermina-

tion of inferred theories, but further knowledge about the variables under investigation (e.g., a variable's arity, distribution type, functional description, etc.) may also be used in selecting plausible theory candidates. Mayo-Wilson (2018) discusses important distributional assumptions that help in disambiguating causal theories, for example when two causal theories (i.e., two causal graphs) are not distinguishable by their conditional independence information, but can be told apart when known facts about marginal distributions are compatible with only one of the graphs. He concludes by hinting at important questions left open.

3.13 Baetu on Multidisciplinary Models and Inter-level Causation

There is a long tradition in philosophy of science and philosophy of mind that understands reality, or at least our description or explanations thereof, in terms of different levels. Different scientific disciplines, despite using vastly different vocabulary to describe phenomena and the world, are not in direct conflict if one understands the targets of their investigations to be separated into distinct and separate strata (Cat, 2017). For example, psychology concerns phenomena at the psychological level, chemistry, the chemical level, and so on. Moreover, levels are ordered by supervenience relations: e.g., the psychological level supervenes on the chemical level. However, Baetu (2018) points out that many causal models used in multidisciplinary scientific investigations in fact relate causal variables at apparently *different* levels. Is this in conflict with the usual supervenience thesis? Baetu (2018) argues rather that these multidisciplinary models should be conceived as level-neutral, rather than multi-level models, for the empirical criteria for inclusion among the models' variable, viz. their susceptibility to effective intervention (Woodward, 2003), are the same. Thus, level-neutrality follows from the epistemic parity of the causal factors, which is a distinct feature of experimental models. Thus, while this doesn't eliminate the philosophical problems about how to relate levels of explanation, it does show that this problem, and issues related to level interaction more generally, are theoretical rather than practical.

As an example, Baetu (2018) considers biopsychosocial models of pain (Asmundson and Wright, 2004; Craig and Versloot, 2011). These models include a variety of physiological, psychological, sociological, and cultural factors determining pain experience as a phenomena that can be the target of medical intervention. Some of these incorporate neural circuit mechanisms, but interacting with higher-order cognitive features as well as factors describing social and cultural determinants of perception and responses to pain, such as (perceptions of) spousal support. Such models are supported by experimental data from brain legion and hypnotic suggestion patients that the mechanisms for sensory perception of pain experience are in fact distinct from the affective aspects of that experience. Nevertheless, these models do not provide complete explanations for pain experience, nor are they intended to; they are rather

effective and useful summaries of the pathways of dependence relevant for effective intervention by medical and psychiatric practitioners.

4 Outlook

Clearly this special issue illustrates that amalgamating evidence in the sciences touches on a variety of philosophical issues concerning confirmation, causation, induction, modeling, experiment, policy, and theory development. Beyond advancing the philosophical discussion of these topics, they also bear upon and deserve further integration with applications in the work of scientists themselves. Already the medical sciences have been a focus here (Wüthrich and Steele, 2018; Mayo-Wilson, 2018), with several examples from cancer research (Wilde and Parkkinen, 2018; Reiss, 2018; Bertolaso and Sterpetti, 2018), pharmaceutical drug trials (Holman, 2018), brain imaging (Danks and Plis, 2018), and biopsychosocial models of pain (Baetu, 2018). However, any science that invites the use of diverse evidence, whether it be sociology (Heesen et al, 2018; Mayo-Wilson, 2018), climate science (Frank, 2018), or even physics (Kao, 2018), can be a target as well, and one should be cautious about extrapolating conclusions valid in one domain to another. Thus, the role of evidence amalgamation in these sciences deserves more attention.

Another application we believe deserves further collaborative attention between scientists, philosophers, and computer scientists is the application of these ideas to big data sets for which considered decisions cannot so easily be made on a case-by-case basis. For example, what is the proper place for (and weight of) expert knowledge in automated assessment of the amalgamated evidence? It seems inevitably necessary in order to formulate the right questions to probe with big data and to keep the answers to those questions computationally tractable, so not all of the process can or should be entirely automated.

Scientists must of course be party to these development, but philosophers are perhaps especially positioned to contribute in a different way, by abstracting general considerations about topics from specific cases—e.g., concerning the confirmatory role that varieties of evidence may and may not play (Claveau and Grenier, 2018), and the justification of procedures for meta-analysis (Vieland and Chang, 2018)—then applying those lessons to new cases in different disciplines. Indeed, what has emerged as a common theme in many contributions to this special issue is how one can and should transfer knowledge from one domain or problem to another: extrapolation from one population to another, the external validity of an experiment, the robustness of a measurement technique, extending static to dynamic causal structure, etc. We enjoin the reader to explore these further.

Acknowledgements This work is supported by the European Research Council (grant 639276) and the Munich Center for Mathematical Philosophy. SCF acknowledges partial support from the European Commission through a Marie Curie International Incoming

Fellowship (PIIF-GA-2013-628533). All guest editors would like to thank Barbara Osimani for her role in brining about this SI as an initial member of the team. We are also grateful to Otávio Bueno for his continued support and wise advice throughout the smooth and enjoyable production process of this SI. Finally, many thanks to all those who submitted manuscripts and the anonymous referees for their time, effort, and excellent work.

References

Asmundson GJG, Wright KD (2004) Biopsychosocial approaches to pain. In: Hadjistavropoulos T, Craig KD (eds) Pain: Psychological Perspectives, Lawrence Erlbaum Associates, Mahwah, NJ, pp 35–58

Baetu TM (2018) On pain experience, multidisciplinary integration and the level-laden conception of science. Synthese

Bertolaso M, Sterpetti F (2018) Evidence amalgamation, plausibility, and cancer research. Synthese

Bohr N (1913) On the constitution of atoms and molecules, part I. Philosophical Magazine 26:1–25

Bovens L, Hartmann S (2002) Bayesian networks and the problem of unreliable instruments. Philosophy of Science 69(1):29–72

Bovens L, Hartmann S (2003) Bayesian Epistemology. Oxford University Press, Oxford

Carnap R (1947) On the application of inductive logic. Philosophy and Phenomenological Research 8(1):133–148

Carnap R (1962) Logical Foundations of Probability, 2nd edn. University of Chicago Press, Chicago

Cartwright N (2004) Causation: One word, many things. Philosophy of Science 71(5):805–819

Cat J (2017) The unity of science. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy, Fall 2017 edn, Metaphysics Research Lab, Stanford University

Cellucci C (2016) Is there a scientific method? The analytic model of science. In: Magnani L, Casadio C (eds) Studies in Applied Philosophy, Epistemology and Rational Ethics volume 25, Springer, Cham, pp 489–505

Cellucci C (2017) Rethinking Knowledge: The Heuristic View. Springer, Cham

Clarke B, Gillies D, Illari P, Russo F, Williamson J (2014) Mechanisms and the evidence hierarchy. Topoi 33(2):339–360

Claveau F (2013) The independence condition in the variety-of-evidence thesis. Philosophy of Science 80(1):94–118

Claveau F, Grenier O (2018) The variety-of-evidence thesis: A Bayesian exploration of its surprising failures. Synthese

Craig KD, Versloot J (2011) Psychosocial perspectives on chronic pain. In: Lynch ME, Craig KD, Peng PWH (eds) Clinical Pain Management: A Practical Guide, Blackwell, Oxford, pp 24–31

Culp S (1995) Objectivity in experimental inquiry: Breaking data-technique circles. Philosophy of Science 62(3):438–458

Cumming G (2012) Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis. Routledge, New York
Danks D, Plis S (2018) Amalgamating evidence of dynamics. Synthese
Debye P, Sommerfeld A (1913) Theorie des lichtelekrischen Effektes vom Standpunkt des Wirkungsquantums. Annalen der Physik 10:873–930
Dowe P (2009) Causal process theories. In: Beebee H, Hitchcock C, Menzies P (eds) The Oxford Handbook of Causation (Oxford Handbooks), Oxford University Press, Oxford, pp 213–233
Du Bois WEB (1996/1899) The Philadelphia Negro: A social study. University of Pennsylvania Press, Philadelphia
Earman J (1992) Bayes or Bust? MIT Press, Cambridge, MA
Einstein A (1907) Die Plancksche Theorie der Strahlung und die Theorie der spezifischen Waerme. Annalen der Physik 22:180–190
Einstein A (1967/1905) On a heuristic point of view about the creation and conversion of light. In: ter Haar D (ed) The Old Quantum Theory, Pergamon, Oxford
Eronen MI (2015) Robustness and reality. Synthese 192(12):3961–3977
Fisher RA (1925) Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh
Frank DM (2018) Ethics of the scientist qua policy advisor: Inductive risk, uncertainty, and catastrophe in climate economics. Synthese
Gagneur J, Friedel C, Heun V, Zimmer R, Rost B (2017) Bioinformatics advances biology and medicine by turning big data troves into knowledge. Informatik-Spektrum 40(2):153–160
Guala F (2010) Extrapolation, analogy, and comparative process tracing. Philosophy of Science 77(5):1070–1082
Hanson NR (1958) Patterns of Discovery. Cambridge University Press, Cambridge
Hanson NR (1960) Is there a logic of scientific discovery? Australasian Journal of Philosophy 38:91–106
Hanson NR (1965) Notes toward a logic of discovery. In: Bernstein RJ (ed) Perspectives on Peirce: Critical Essays on Charles Sanders Peirce, Yale University Press, New Haven, CT, pp 42–65
Heesen R, Bright LK, Zucker A (2018) Vindicating methodological triangulation. Synthese
Holman B (2018) In defense of meta-analysis. Synthese
Holman B, Bruner J (2017) Experimentation by industrial selection. Philosophy of Science 84(5):1008–1019
Horwich P (1982) Probability and Evidence. Cambridge University Press, Cambridge
Hüffmeier J, Mazei J, Schultze T (2016) Reconceptualizing replication as a sequence of different studies: A replication typology. Journal of Experimental Social Psychology 66:81–92
Hunter A, Williams M (2012) Aggregating evidence about the positive and negative effects of treatments. Artificial Intelligence in Medicine 56(3):173–190

Jones N (2018) Inference to the more robust explanation. The British Journal for the Philosophy of Science 69(1):75–102
Jukola S (2015) Meta-analysis, ideals of objectivity, and the reliability of medical knowledge. Science & Technology Studies 28(3):102–121
Jukola S (2017) On ideals of objectivity, judgments, and bias in medical research – A comment on Stegenga. Studies in History and Philosophy of Biological and Biomedical Sciences 62(Supplement C):35–41
Kao M (2018) Unification beyond justification: A strategy for theory development. Synthese
Keynes JM (1921) A Treatise on Probability. MacMillan, London
Kilinc B (2012) Meta-analysis as judgment aggregation. In: de Regt HW, Hartmann S, Okasha S (eds) EPSA Philosophy of Science: Amsterdam 2009, Springer Netherlands, Dordrecht, pp 123–135
Kuorikoski J, Marchionni C (2016) Evidential diversity and the triangulation of phenomena. Philosophy of Science 83(2):227–247
Kuorikoski J, Lehtinen A, Marchionni C (2012) Robustness analysis disclaimer: please read the manual before use! Biology & Philosophy 27(6):891–902
Lagoa CM, Barmish B (2002) Distributionally robust Monte Carlo simulation: A tutorial survey. IFAC Proceedings Volumes 35(1):151–162
Landes J (2018) Variety of evidence. Erkenntnis Forthcoming
Landes J, Osimani B (2019) Varieties of error and varieties of evidence. Philosophy of Science Forthcoming
Laudan L (1980) Why was the logic of discovery abandoned? In: Nickles T (ed) Scientific Discovery, Logic, and Rationality, D. Reidel, Dordrecht, pp 173–183
Lehtinen A (2018) Derivational robustness and indirect confirmation. Erkenntnis Forthcoming
Leuridan B, Weber E (2011) The IARC and mechanistic evidence. In: Illari P, Russo F, Williamson J (eds) Causality in the Sciences, Oxford University Press, Oxford, pp 9–109
List C, Goodin RE (2001) Epistemic democracy: Generalizing the Condorcet jury theorem. Journal of Political Philosophy 9(3):277–306
Lloyd EA (2015) Model robustness as a confirmatory virtue: The case of climate science. Studies in History and Philosophy of Science 49:58–68
Magnani L (2001) Abduction, Reason, and Science: Processes of Discovery and Explanation. Plenum, New York
Magnani L (2009) Creative abduction and hypothesis withdrawal. In: Meheus J, Nickles T (eds) Models of Discovery and Creativity, Springer Netherlands, Dordrecht, pp 95–126
Mayo-Wilson C (2011) The problem of piecemeal induction. Philosophy of Science 78(5):864–874
Mayo-Wilson C (2014) The limits of piecemeal causal inference. British Journal for the Philosophy of Science 65(2):213–249
Mayo-Wilson C (2018) Causal identifiability and piecemeal experimentation. Synthese

Mebius A, Kennedy AG, Howick J (2016) Research gaps in the philosophy of evidencebased medicine. Philosophy Compass 11(11):757–771

Pearl J (2000) Causality: Models, Reasoning, and Inference, 1st edn. Cambridge University Press

Planck M (1967/1900) On the theory of the energy distribution law in the normal spectrum. In: ter Haar D (ed) The Old Quantum Theory, Pergamon, Oxford

Poellinger R (2018) Analogy-based inference patterns in pharmacological research. In: La Caze A, Osimani B (eds) Uncertainty in Pharmacology: Epistemology, Methods, and Decisions, Boston Studies in Philosophy of Science, Springer, (forthcoming)

Raerinne J (2013) Robustness and sensitivity of biological models. Philosophical Studies 166(2):285–303

Reichenbach H (1971) The Direction of Time. University of California Press, Berkeley

Reiss J (2015) A pragmatist theory of evidence. Philosophy of Science 82(3):341–362

Reiss J (2018) Against external validity. Synthese

Reiss J, Ankeny RA (2016) Philosophy of medicine. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy, Summer 2016 edn, Metaphysics Research Lab, Stanford University

Romero F (2016) Can the behavioral sciences self-correct? A social epistemic study. Studies in History and Philosophy of Science 60:55–69

Rubinstein RY, Kroese DP (2016) Simulation and the Monte Carlo method, 3rd edn. John Wiley & Sons, Hoboken, NJ

Russell B (1912) On the notion of cause. Proceedings of the Aristotelian Society 13:1–26

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

Salmon W (1984) Scientific Explanation and the Causal Structure of the World. Princeton University Press, Princeton

Schaffner K (1993) Discovery and Explanation in Biology and Medicine. University of Chicago Press, Chicago

Schickore J (2014) Scientific discovery. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy, Spring 2014 edn, Metaphysics Research Lab, Stanford University

Schupbach JN (2018) Robustness analysis as explanatory reasoning. British Journal for the Philosophy of Science 69(1):275–300

Spirtes P, Glymour C, Scheines R (2000) Causation, Prediction, and Search. Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA

Staley KW (2004) Robust evidence and secure evidence claims. Philosophy of Science 71(4):467–488

Steel D (2008) Across the Boundaries: Extrapolation in Biology and Social Sciences. Oxford University Press, Oxford

Steel D (2014) Philosophy and the Precautionary Principle: Science, Evidence, and Environmental Policy. Cambridge University Press, Cambridge

Stegenga J (2011) Is meta-analysis the platinum standard of evidence? Studies in History and Philosophy of Biological and Biomedical Sciences 42(4):497–507

Stegenga J, Menon T (2017) Robustness and independent evidence. Philosophy of Science 84(3):414–435

Suppes P (1970) A Probabilistic Theory of Causality. North-Holland, Amsterdam

Sutton AJ, Higgins JPT (2008) Recent developments in meta-analysis. Statistics in Medicine 27(5):625–650

Sutton AJ, Abrams KR, Jones DR (2001) An illustrated guide to the methods of meta-analysis. Journal of Evaluation in Clinical Practice 7(2):135–148

Vieland VJ, Chang H (2018) No evidence amalgamation without evidence measurement. Synthese

Weisberg M (2006) Robustness analysis. Philosophy of Science 73(5):730–742

Weitzman ML (2009) On modeling and interpreting the economics of catastrophic climate change. Review of Economics and Statistics 91(1):1–19

Whewell W (1840) The Philosophy of the Inductive Sciences. John W. Parker, London

Whewell W (1858) Novum Organon Renovatum. John W. Parker, London

Wilde M, Parkkinen VP (2018) Extrapolation and the Russo-Williamson thesis. Synthese

Woodward J (2003) Making Things Happen: A Theory of Causal Explanation. Oxford University Press, Oxford

Woodward J (2006) Some varieties of robustness. Journal of Economic Methodology 13(2):219–240

Worrall J (2007) Evidence in medicine and evidence-based medicine. Philosophy Compass 2(6):981–1022

Wüthrich N, Steele K (2018) The problem of evaluating automated large-scale evidence aggregators. Synthese