Analogy-Based Inference Patterns in Pharmacological Research*

Roland Poellinger

Abstract Analogical arguments are ubiquitous vehicles of knowledge transfer in science and medicine. This paper outlines a Bayesian evidence-amalgamation framework for the purpose of formally exploring different analogy-based inference patterns with respect to their justification in pharmacological risk assessment. By relating formal explications of similarity, analogy, and analog simulation, three sources of confirmatory support for a causal hypothesis are distinguished in reconstruction: relevant studies, established causal knowledge, and computational models.

Key words: scientific inference, pharmacology, epistemology, Bayesian confirmation, evidence, relevance, similarity, analogy, computer simulation

1 Introduction: scientific inference in pharmacology

Pharmacological research is often driven by many forces at once: Cost effectiveness must be balanced against extensive data-collecting, potential risk against probable benefit, and breadth of applicability against well-documented higher confidence for smaller target groups. Many such decisions must be taken during the development stages of a certain drug before the desired effectiveness and safety level is reached and the drug is allowed to be marketed. A language for expressing both benefit and safety is found in the probabilistic language of expected utilities and dis-utilities. Nevertheless, the formalization of a given decision problem in such vocabulary can only be as informative about future drug users as the evidence it is rooted in. Yet,

Roland Poellinger
Munich Center for Mathematical Philosophy, LMU Munich
e-mail: r.poellinger@lmu.de

* This work is supported by the European Research Council (grant 639276) and the Munich Center for Mathematical Philosophy (MCMP).

whether it is all the evidence about the drug's effects that is to be taken into consideration, or only the best evidence available, is the subject of an ongoing discussion in the philosophy of medicine. A recent paper by Landes, Osimani, and Poellinger (Landes et al, 2017) explores the possibility of amalgamating all available evidence in a Bayesian reconstruction of scientific inference for the integrated probabilistic assessment of a drug's causal (side-)effects. Key to this endeavor is the distinction of the conceptual levels involved: (i) the causal hypothesis, (ii) testable indicators of the causal claim, i.e., theoretical consequences of the causal hypothesis, and (iii) concrete evidence (actual data) speaking for or against the respective indicators (thereby indirectly supporting the hypothesis, or not). Close to pharmacological practice, it is furthermore useful to introduce an additional *meta-evidential* level for the purpose of encoding the qualified assessment of the data at hand. This allows to conceptually distinguish the *content* of an evidential report from its *weight* in the evaluation of the hypothesis: Different pieces of evidence may possess different *levels of significance* (iv).

One important justification of the confirmatory support a piece of evidence lends to a given hypothesis (by virtue of it being evidence for an indicator of the very hypothesis) is the postulate (or implicit assumption) of analogy between the circumstances generating the evidence and the hypothesis' intended (future) scope of application. Sir Austin Bradford Hill lists analogy as one of his famous guidelines towards an informed assessment of potential causes in epidemiology:

> In some circumstances it would be fair to judge by analogy. With the effects of thalidomide and rubella before us we would surely be ready to accept slighter but similar evidence with another drug or another viral disease in pregnancy.[2] (Hill, 1965, p. 11)

The aim of this paper is to explore different analogy-based inference patterns in the above-sketched layered reconstruction of scientific reasoning, with respect to their justification in pharmacological risk assessment. In particular, different aspects of similarity shall be made transparent in order to compare conceptually different types of evidence in analogical reasoning. Three interrelated questions shall be addressed in the following sections:

1. How can analogy considerations be used to explicate the relevance of evidence for a hypothesis under consideration? (Sec. 2)
2. How can an already well-established hypothesis be used as a supporting analog in confirming a hypothesis about a similar drug? (Sec. 3)
3. By what standards can a mechanistic computational model of a substance's causal effects as a theoretical analog lend evidential support to a causal hypothesis about the actual substance? (Sec. 4)

[2] In this passage, Hill refers to (i) severe disabilities (even death) among babies linked to the over-the-counter drug thalidomide, prescribed in the 1960s in Germany as Contergan to alleviate morning sickness in pregnant women, and to (ii) miscarriage or children born with the congenital rubella syndrome (CRS) due to infection by the rubella virus during pregnancy.

Before we can trace the role of analogy in pharmacological research, though, a formal reconstruction of the dynamics underlying scientific hypothesis testing shall be outlined in the following.

1.1 Evidence amalgamation and hypothesis confirmation

Causal inference in pharmacology is a difficult task due to sometimes sparse, oftentimes very heterogeneous evidence, different standards for evidence evaluation and integration, and many sources of random and systematic error. In their pluralist, conciliatory approach, Landes, Osimani, and Poellinger (Landes et al, 2017) propose a blueprint for tracing the epistemological dynamics of evidence amalgamation, viewing risk assessment in pharmacology from a meta-perspective. In their framework, causal hypotheses (about adverse drug reactions) and evidential reports (about concrete studies) are related in such a way, that successively cumulating evidence allows for probabilistic causal inference. One important point of departure for this approach is Bovens' and Hartmann's general reconstruction of scientific inference (Bovens and Hartmann, 2003) in which the epistemic dynamics of all the dependencies between a hypothesis, testable indicators implied by the hypothesis, and evidence for/against such indicators can be visually traced in the graph of a Bayesian network: The conceptual categories are depicted as layers of nodes, with directed edges between the layers marking those paths along which confidence in the hypothesis is boosted (or lowered). The probabilistic model underneath the graphical representation supplements the Bayesian network with quantitative information by encoding (i) conditional degrees of (un)certainty about all variables, and (ii) how these degrees will change under local updates.

The example in Fig. 1 illustrates the epistemological structure: The causal hypothesis (Hyp) entails n theoretical indicators (i.e., testable consequences $Ind_1, \ldots, Ind_n$), which are to be supported by concrete evidence reports $Rep_1, \ldots, Rep_n$ on the lowest level. Since evidence reports are based on concrete data (e.g., clinical trials, historical studies, or lab experiments), it will be desirable in many cases to modulate their significance for the assessment of a particular causal hypothesis for various reasons: The source of information seems undependable, the quality of a study might be doubtful, the reliability of method or measurement device is not guaranteed, test group and target group deviate in relevant details, and so on. To be able to express such levels of significance, the report nodes in the graph come with an additional weight node α (for the moment simply a blackbox placeholder for the aforementioned significance dimensions, used below to encode *evidential relevance* in particular, see Sec. 2). The degree to which a given evidence report supports the hypothesis depends precisely on its "weight", i.e., its relevance to the hypothesis, the reliability of the source, the error-proneness of the methods used to generate the data, and so on. Such weight nodes might in general be shared between reports – for example in cases where more than one report is based on data generated by the same

measurement device. While Bovens and Hartmann utilize this weighting parameter to explicate the reliability of a report, Landes, Osimani, and Poellinger – aiming at hypothesis confirmation in pharmacology – split this weighting parameter into two variables (*reliability* and *relevance*) in order to distinguish the quality of method and information source from questions of external validity (by encoding the degree to which study results can be *extrapolated* to the target).[3]

What it means to be an observable consequence of the hypothesis is implicitly stated by the following inequality (see also Bovens and Hartmann 2003, p. 90):

$$P(Ind_i \,|\, Hyp) = p_i > q_i = P(Ind_i \,|\, \neg Hyp), \tag{1}$$

where 'Hyp' is to be understood as shorthand for '$Hyp = \text{true}$', and analogously for the other variables (in the following, context disambiguates if a variable or its instantiation is referred to).

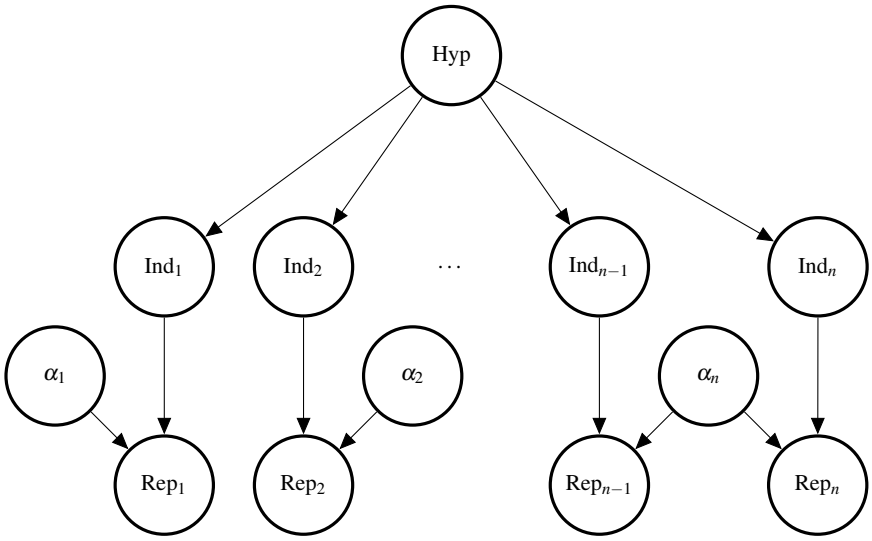

Fig. 1: Example of how a causal hypothesis, multiple testable consequences, and weighted evidence reports might be related in a Bayesian network.

[3] As implied by this way of distinguishing reliability and relevance, the relevance weight (attached to a given evidential report) is really meant here to capture the degree of external validity. Of course, there are other ways in which a study can be relevant to the hypothesis – for example, if it is conducted by an acknowledged authority. In the framework of Landes, Osimani, and Poellinger (Landes et al, 2017), this way of being relevant to the hypothesis would be encoded in the reliability weight, which collects all sources of bias.

The conditional in/dependencies between the epistemic categories can be read off the structure in Fig. 1 by utilizing the graphical *d*-separation criterion:[4] All directed edges mark positive influence between adjacent nodes (variables, respectively), which makes all *Hyp*, *Ind*, and *Rep* variables positively correlated (conditional on the empty set). Following Bovens and Hartmann (2003), the weight variables are *independent* of the theoretical categories (i.e., hypothesis and indicators) and may be used to account for exceptions.[5] I will treat the α-weights as *binary modulators* throughout the paper (with $\alpha = 1$ indicating *significance* and $\alpha = 0$ indicating *insignificance*, respectively); nevertheless, the formal framework can easily be extended to allow for a more fine-grained (in many cases more appropriate) representation.

The in/dependencies visualized in Fig. 1 can be summed up probabilistically as follows, for all i ($1 \leq i \leq n$):

$$Hyp, Ind_i \perp\!\!\!\perp \alpha_i \tag{2}$$

$$P(Rep_i \mid \ Ind_i, \alpha_i = 1) = 1 \tag{3}$$

$$P(Rep_i \mid \neg Ind_i, \alpha_i = 1) = 0 \tag{4}$$

$$P(Rep_i \mid \ Ind_i, \alpha_i = 0) = \tag{5}$$

$$P(Rep_i \mid \neg Ind_i, \alpha_i = 0) =$$

$$P(Rep_i \mid \quad \alpha_i = 0) = q_i$$

Eq. 2 precisely encodes the independent assignment of the weighting parameter. Eq. 3 and Eq. 4 show that an evidence report marked as "significant" to the hypothesis is aligned with the respective indicator. The loose term "significance" is meant to be understood in this context as a quite general concept up for interpretation – it does not refer to statistical significance of study results, but rather to the quality of evidence (as meta-information), e.g., in terms of *reliability* or *relevance*, as discussed below. Eq. 5 explicates what happens when an evidence report is qualified as "utterly insignificant" to the hypothesis: In that case, whether the report speaks for or against the respective indicator becomes probabilistically independent of the truth value of this very indicator variable (with $0 \leq q_i \leq 1$). In particular, an evidence report considered *irrelevant* for the investigated hypothesis should not influence one's

[4] The graphical *d*-separation criterion (with *d* for *directional*) distinguishes conditionally dependent (sets of) variables from conditionally independent ones by drawing on structural information, i.e., on how arrows are directed along the paths between the (sets of) variables under consideration; see, e.g., Geiger et al (1990).

[5] In this case, independence between the weight variables and the hypothesis "may or may not be a realistic assumption", as Bovens and Hartmann concede, and they extend their discussion to cases where such weight nodes (reliability of evidence reports) and the hypothesis are made dependent through auxiliary theories (see Bovens and Hartmann 2003, pp. 107ff.). For the purpose of this paper, though, the standard for assigning values to weight variables is assumed to be fixed prior to hypothesis testing.

confidence in the truth of the hypothesis' indicators.[6]

With this Bayesian framework at hand, explicating what it means for a given piece of "significant" evidence to confirm a causal hypothesis is straightforward: An evidence report Rep_k is said to confirm the hypothesis Hyp iff it raises Hyp's probability, i.e., iff

$$P(Hyp\,|\,Rep_k) > P(Hyp). \tag{6}$$

Obviously, this inequality does not hold for $\alpha_k = 0$, i.e., for "insignificant" evidence reports (unreliable sources of evidence or evidence irrelevant for the hypothesis etc.). Consequently, what it means for a given piece of evidence to be "significant" to a hypothesis must be defined before we can start confirming this very hypothesis. Analogical reasoning shall be exploited towards that goal in the following.

1.2 Analogy as inferential pattern

At the heart of analogical arguments lies the assumption of (sufficient) similarity between the two relata (or aspects thereof, respectively). The concept of similarity comes with a mixed bag of notorious epistemological issues of its own, though. For example, if two drugs are to be related in an analogical argument, the following questions come to mind: What does it mean to be sufficiently similar in the case under consideration? In what way does the difference between the first and the second drug influence changes in expected outcome values? How specific are a drug's properties? If they are highly specific – to what extent can this drug be used in an analogical argument, if at all? Despite these conceptual difficulties, science and history are full of successful examples of analogical reasoning, even in the case of scientific discovery: In the nineteenth century, secured knowledge about acoustics was employed in the discovery of spectral lines. Guided by the image of a harmonic oscillator, physicists were able to focus their attention to groups of spectral lines with specific frequency patterns from the beginning (see Bartha's in-depth overview of analogical arguments in Bartha 2013, as well as Unruh 2008, Dardashti et al 2017, and Hesse 1952 for discussions of analog arguments in physics).

In pharmacology and epidemiology, reasoning by analogy is a key mode of knowledge transfer from study to target population. Indeed, because of the context sensitivity of many causal associations in the biological realm, these associations can hold only in specific populations, and therefore evidence about causal effects related to one population may not license similar conclusions about another population, unless the two populations have been established as *analogous* (in relevant respects). Knowledge about an agent's mechanisms and about its impact on the biological environment might be sparse and come from quite heterogeneous sources,

[6] The Bayes net structure in Fig. 1 for example illustrates that Ind_1 is influenced by (since d-connected to) α_1 once we know the value of the "collider variable" Rep_1.

however. Yet, already if only little information about the agent's class of molecules is available, for example, this can justifiably be exploited for causal assessment *via analogy*.

As Bartha (2010) points out, analogical arguments fall into a different category than regular evidence as such. Consider the general form of an analogical argument (see Bartha 2013):

1. y is similar to x (in certain known respects), **(A)**
2. x has additional property A,
3. therefore: y has property A' (similar to A).

This argument obviously encodes basic evidence about x which is then used to make inferences about y. It is unclear, as Bartha points out, how such a structured argument could be encapsulated in a single *evidential proposition* to figure as a condition in the scheme of Bayesian confirmation, stated in Eq. 6 above. Moreover, according to Bartha, even if we were to condense a full analogical argument into one such evidential proposition A, the postulated similarity between source and target domain must have been established *before* A can be used in confirmation, which makes A "old evidence" and useless for confirmation (since it does not boost the degree of belief in the hypothesis). This becomes obvious once Eq. 6 is relativized to fixed background knowledge K. With the analogical argument A entailed by K, we get $P(Hyp\,|\,A,K) = P(Hyp\,|\,K)$.

With the above-sketched Bayesian reconstruction of scientific hypothesis testing at hand, we already have a structure-rich framework utilizable for a different formalization of analogical reasoning: The key idea is to understand analogy not as single nodes, but to trace confirmation by analogy along the edges in the graph. In other words, analogical arguments shall be understood and expressed rather as inferential patterns than as "evidence nodes".

The evidence-amalgamation framework can be operationalized for this purpose in the following ways:

1. The question of applicability of a study's findings is best phrased in terms of correspondence between study and target populations, where correspondence shall be understood as *similarity* (above a certain threshold): If study and target populations are *sufficiently similar*, researchers are licensed to reason about causal links in the target population by analogy with their test cases. The extent to which this kind of transfer is licensed is encoded in the framework as *relevance* of available reports (such that independent reports are each assigned a different degree of relevance with respect to the investigated hypothesis). The more characteristics are shared between study and target, the higher the relevance of evidence obtained in this study for the hypothesis under investigation. This case is treated in Sec. 2.

2. The investigated causal hypothesis Hyp may be related to a second causal hypothesis Hyp^* which has already been established in previous studies. Now, if scientists have sufficient reason to postulate analogy between Hyp and Hyp^* (e.g., a high degree of chemical or functional similarity), knowledge about this second causal hypothesis supports the first hypothesis Hyp *via analogy*. This case requires an extension of the layered network introduced above and will be treated in Sec. 3.
3. As a special sub-case of 2., I will discuss how Hyp can possibly be confirmed by way of computational modeling (instead of by relating it to a second causal hypothesis Hyp^* about a biological system): In cases where scientists want to confirm mechanistic assumptions, computer simulation has become a viable alternative to costly, unethical, or otherwise inaccessible experimental studies. Assumptions about the efficacious mechanistic relationships may be modeled in a *virtual analog* to find out more about the protein a drug binds to or to simulate an agent's interaction with the biological system, for instance, w.r.t. dosage (see, e.g., Britton et al 2013 and Carusi et al 2012). In order to treat this special case I will tweak the network structure even further in Sec. 4.

My aim in this paper is to exploit the concept of analogy in explicating these different knowledge transfer strategies. To this end, I will investigate the confirmatory dynamics of analogical reasoning in a formal way: All three abovementioned cases shall successively be located in the evidence-amalgamating framework for the purpose of unifying the inferential patterns and emphasizing their structural differences at the same time. I will start by discussing ways of designing possible (comparative or numerical) distance measures for expressing similarity (between substances, populations, etc.).

2 Learning from relevant evidence

2.1 Heterogeneous evidence

The evidence-amalgamating framework, as introduced above, presents itself as a tool for describing specific cases in that it can be adapted to accommodate specific pieces of evidence for (or against) specific theoretical consequences of the causal hypothesis under consideration. At the same time, it is meant to be understood as a normative statement about the quality of the causal assessment: The more information about the indicator variables is available, the more reliable the assessment of the hypothesis. This is especially true in the case of *causal* hypotheses, given the variety of methodological approaches towards discovering causal associations in the sciences.[7]

[7] Landes et al (2017) contains a non-exclusive list of six causal indicators derived from Hill's guidelines in Hill (1965). See also Poellinger (n.d.) for a discussion of the conceptual relationships of these causal indicators and the ramifications of theory choice in causal assessment.

The multi-layered framework introduces report nodes as placeholders for heterogeneous pieces of evidence. The reports can come from all levels of the "evidence hierarchy" – the framework treats all these levels as relevant in causal assessment: Systematic reviews, meta-analyses of randomized clinical trials or observational studies, comparative non-randomized studies (cohort or case-control studies), evidence of (sub-)mechanisms from basic science (lab experiments etc.), as well as expert judgment. Relevant information may come also from single case reports, case series, and animal studies. Amalgamating these different types of evidence has not only epistemological value (since it traces the epistemological dynamics from observed signals of Nature to the establishment of hypotheses) but also methodological value (since it readily accommodates different approaches towards evidence assessment and related disputes like the seeming tension between paradigms such as best-evidence vs. pluralistic approaches).

2.2 Relevance

However heterogeneous the evidence, different evidence reports (understood as data interpreted relative to the hypothesis under consideration) will be attributed different levels of significance for the hypothesis: Virtually no pharmacological study allows for straightforwardly transferring shown results from studied population to target population of possible future drug users – population size, inherent structure, specific circumstances, possible interactions with uncontrolled substances, etc. might be similar, but will virtually never be the exact same (see Chan and Altman 2005, p. 1180, Doll and Peto 1980, Worrall 2007, p. 992). Furthermore, patient inclusion criteria for participation in RCTs will skew the inference transfer from study to target population even more (see, e.g., Revicki and Frank 1999 and Upshur 1995, p. 483). When the results of animal studies are to be applied to a population of future human drug users it becomes quite clear that questions of applicability must be settled first before such transfer is licensed.[8] Moreover, in their recent analysis of the role of evidence about a substance's mechanisms in risk assessment, Luján, Todt, and Bengoetxea (Luján et al, 2016) point towards the lack of guarantee that similarity of modes of action may warrant extrapolation of phenotypic effects from one chemical to another – chemicals considered similar in important respects might not necessarily produce similar effects (for a given population). And finally, when laboratory experiments yield information about some component of the causal mechanism (e.g., at the molecular level or in terms of cell behavior), the significance of this (partial) result for the hypothesis about the entire mechanism must be determined first.[9]

[8] LaFollette and Shanks (1995) argue, e.g., that animal studies are limited to hypothesis *generation*.

[9] In particular, the question must be answered whether the partial result can be combined in an additive fashion with information about further sub-mechanisms, or whether complex interdependencies forbid partitioning the full mechanism into stand-alone modules.

In all these cases scheme (A) is applied implicitly or explicitly: The hypothesis is confirmed by analogy, i.e., its degree of confidence is raised by a given evidence report, once similarity between studied subjects (or objects) and intended application is established. In the framework of Landes, Osimani, and Poellinger, this type of significance of a given evidence report for a hypothesis to be tested is expressed as evidential relevance factor (whose value is to be determined outside the model), cf. Landes et al 2017, Sec. 3.3:[10]

> Ideally, pharmacological studies would license the same inferences for the studied population and the target population of future drug users. In reality, studies are not conducted on the *entire population of future drug users* but on a much smaller number of patients, see Chan and Altman 2005, p. 1180, Button et al 2013 for this problem in neuro science, Doll and Peto 1980 in cancer research and a philosophical discussion of this problem in Worrall 2007, p. 992. Additionally, studied populations, in particular in RCTs, often fail to be representative for the target population due to strict patient inclusion criteria, see Revicki and Frank 1999 and Upshur 1995, p. 483. Therefore, there is a need to reason by analogy from the studied population to the population of interest [...] The relevance pertaining to an item of evidence measures how well the observed results in a study population can be transferred to the target population of future drug users.

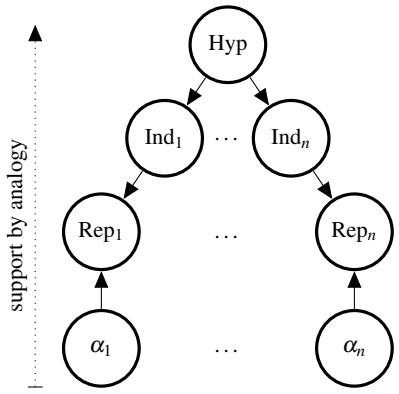

Fig. 2: Support by analogy from the level of evidence reports ($Rep_1, \ldots, Rep_n$) upwards to the hypothesis under consideration.

Fig. 2 illustrates this relationship between an item of evidence and the target: When hypothesis Hyp is on the test bench, testable consequences $Ind_1, \ldots, Ind_n$ (indicators of causation) are supported (or not) by evidence reports $Rep_1, \ldots, Rep_n$. Since the framework is set up in Bayesian fashion, once the value of Rep_k is fixed

[10] Note that I am distinguishing *evidential* relevance (as a property of evidence relevant in causal assessment) from *causal* relevance (as a property of a variable causally relevant to a second variable in a causal model), cf. footnote 18. Also see footnote 3 above for a remark on other interpretations of relevance.

(e.g., to $Rep_k = true$, indicating that there is evidence in support of $Ind_k = true$), the respective significance variable α_k will determine the relative strength of Rep_k's contribution to Hyp (since – in Bayesian terms – knowing a collider's value opens the flow of information along the path the collider lies on). Note, that – in principle – two (or more) report nodes might share the same significance, as exemplified in Fig. 1, when inter-dependencies are to be indicated (in terms of either *reliability* or *extrapolability*, i.e., *relevance*).

If we now interpret the α-layer as encoding information about the relevance of our evidence reports, we can try to make the relationship between α_k and Hyp transparent in terms of similarity.[11]

2.3 Measuring distance

In order to be able to compare the hypothesis and incoming evidence reports, the structural ingredients of these categories shall be made explicit in more detail. In risk assessment, Hyp is meant to be a causal hypothesis about a specific (harmful) side-effect E of a certain drug D – to be abbreviated as 'D ⓒ E' in the following. Since the evidence-amalgamating framework is intended to yield quantitative (probabilistic) information for decision-making (e.g., in terms of expected utilities), D ⓒ E must be evaluated in a quantitative framework that allows for the specification of all relevant causal relations in a suitable formal model M. One important modeling decision will be the separation of parameters influenced *within* the model (i.e., 'observed' or 'endogenous' variables – in particular, the modeled effects and side-effects of interest, along with further variables on the causal path from drug to potentially harmful outcome) from *ceteris paribus conditions* ('context' or 'world' or 'circumstances', encoded as 'exogenous' variables in a causal model). A fourth parameter shall be added to our language to express precisely these contextual conditions, i.e., the ceteris-paribus aspects of the population for which the causal mechanism is to be tested. The parameter U shall sum up those characteristics of population plus environment that will not be changed by, e.g., a clinical trial (as, e.g., average age, nutrition habits, aspects of the medical history, socio-economic status, etc.).[12] We can accordingly expand Hyp to encode the aforementioned ingredients:

[11] Note that for such an interpretation the prior of the network is required to be set up in such a way that the αs render the *Rep*s independent of (i.e., irrelevant to) the hypothesis Hyp in the extreme case. Formally: $P(Hyp \,|\, Rep_k = true, \alpha_k = irrelevant) = P(Hyp \,|\, \alpha_k = irrelevant) = P(Hyp)$. See also my discussion of the prior in Sec. 1.1 above.

[12] This suggests that the distinction between changeable and unchanged aspects of the study population will be a static one prior to modeling. Nevertheless, Paul and Healy discuss cases in which the modeler is forced to revisit her model because a clinical trial impacts on relevant characteristics of the population in a sort of feedback loop (see Paul and Healy 2016 on *Transformative Treatments*). For the present purpose it is uncritical to assume that the initial model can be refined at later stages to accommodate previously exogenous assumptions into the model as endogenous parameters/relations.

$$Hyp: \; D \copyright E \text{ in model } M \text{ with context } U \tag{7}$$

Explicating the causal hypothesis in this way is not meant here to express a specific commitment as to how © is to be interpreted. In particular, it is not meant to confine the investigation to an epistemic, model-relative theory of causality. This specific way of encoding the causal hypothesis is rather meant to be as informative as possible about the intended scope of the causal claim and to mark deviations between studies and target with respect to population characteristics (information about age, multimorbidity, and so on, encoded in $U_{(k)}$) and structural assumptions (confounders, causal history, possible disablers, etc., encoded in $M_{(k)}$).[13]

Now, using a piece of evidence for the confirmation of a hypothetical causal effect E *by analogy* may be licensed for different reasons:

1. Similar substance D: The substance used on the study population is similar to the substance to be applied to the target population. Of course, if an experimental setup is designed with $D \copyright E$ in mind, the experiment will generally test the causal effects of D itself. Nevertheless, what may vary is the administered dosage or indeed the final formula with respect to additional components.
2. Similar causal model M: Causal efficacy is judged with respect to similar modeling assumptions (variables and their order, causal in/dependencies, and so on) in the study and for the intended application. For example, when cohort studies point to the existence of a *causal path* between paracetamol and asthma, knowledge about these studies can only be transferred to a population of future drug users if the core causal assumptions are kept fixed. Postulating a deviating causal structure for the target, e.g., a *common cause* of paracetamol use and asthma instead of a causal path, would prevent using evidence from such cohort studies for predicting the effects of certain policy interventions towards decreasing asthma morbidity.
3. Similar context U: Contexts, i.e., study and target population plus respective environment, are similar in relevant respects. For example, when drugs are tested on animals first, pigs are one of the preferred species, because they share much of the human genetic make-up and consequently many of the complex genetic diseases, such that a drug's causal effects in pigs can to some extent be used in inference about human biology.

Scheme (A) above can now be adjusted for *learning from relevant evidence* in the evidence-amalgamating layered reconstruction of scientific inference:

1. $\langle D, M, U \rangle$ is similar to $\langle D_k, M_k, U_k \rangle$, **(A*)**
   therefore, by stipulation: Rep_k is relevant for Hyp, i.e., α_k is high,

[13] If the causal hypothesis is thought of as a causal graph, D, E, and U are meant to represent designated (sets of) variables with token values in a causally interpreted structure M (possibly encoding the specifics of direct causal relations and assumptions about causal in/dependencies on type level). Note that, more generally, M and M_k can be thought of as *sets of structural constraints*, i.e., as *classes of causal graphs*.

2. Rep_k: $\Phi[D_k, E_k, M_k, U_k]$
   with E_k similar to the hypothesized E,
3. therefore: $\Phi'[D, E, M, U]$
   with Φ' similar to Φ,

where Φ is a quantitative or qualitative statement about (the components of) the study (and Φ' a statement about the target, respectively). For example, Φ might encode the conditional probability $P(E_k \,|\, D_k) = q$; therefore, *by analogy*, Φ': $P(E \,|\, D) \approx q$ (for high α_k). Note that, by design, Rep_k will influence Hyp on a path through a suitable (set of) mediating indicator variable(s). If now Rep_k makes a positive contribution to the indicator level, it will boost confidence in Hyp, thereby confirming $D \copyright E$:

$$P(D \copyright E \,|\, Rep_k, \alpha_k) \geq P(D \copyright E).$$

The inference pattern (A*) crucially relies on a high relevance parameter α_k which is determined in assessing the *similarity* of the triples $\langle D, M, U \rangle$ and $\langle D_k, M_k, U_k \rangle$. The literature on similarity conceptualizations and measures is vast, spanning from Lewis' lexicographic *comparative similarity* of possible worlds (Lewis, 1973a) to Shepard's *geometric account* in terms of vector distances (Shepard, 1980) and to *contrast approaches* evaluating similarity by weighting and comparing properties (see, e.g., Tversky 1977, Weisberg 2012, and Weisberg 2013).[14] Some are formulated in a symmetric way, others encode asymmetric relations. All approaches seem tied to specific practices and purposes with seemingly none universally applicable. Nevertheless, with analogical arguments employed successfully in pharmacological research, general features of the similarity concept needed for (A*) shall be outlined in the following.

Claiming that $\langle D, M, U \rangle$ is similar to $\langle D_k, M_k, U_k \rangle$ (as in line 1 of the inference scheme for relevant evidence) quite generally means that the individual components are in the same equivalence class (by pairs), formally: $D_k \in [D]_\sim$, $M_k \in [M]_\sim$, and

[14] These accounts of similarity share the intuition that comparing two things means (i) comparing certain aspects of those things and (ii) aggregating one's evaluation of those aspects in a certain manner.
Lewis' idea of comparative similarity is tightly connected to his concept of causation, where a cause–effect relation is evaluated in terms of the corresponding counterfactuals. The cause reveals its power in the effect event where the rest of the world remains unperturbed, i.e., as *similar as possible* to the state of world prior to the cause event. Lewis suggests a priority ordering for the assessment of similarity, where local changes in physical facts are understood as a lesser deviation from actuality than far-reaching global changes in natural laws, see Lewis (1973b).
The geometric account locates an object's properties (deemed relevant for comparison) in a multi-dimensional space by assigning a specific value to each of those properties. Similarity is then spelled out in terms of vector distance from a reference object.
The question of how to assign such values is circumvented in the contrast approach which deals well with similarity as *partial identity*, since in this approach degrees of similarity are assessed by assigning weights to *co-instantiated identical properties* (which might make the approach suitable rather for comparing different states of one and the same object than for comparing different objects).

$U_k \in [U]_\sim$.[15] These equivalence classes are induced by (i) the components of the hypothesis, D, M, U, and (ii) highly purpose-oriented considerations as to balancing the properties to be compared (i.e., as to how properties are to be ordered in Lewis-style approaches, as to how which properties are metrized in geometric approaches, and as to how properties are to be weighed in contrast comparisons). Presenting similarity this way allows a research community to fill in a meaningful, informative understanding of similarity, and it separates purpose-driven considerations concerning how to suitably define an equivalence relation from the decision about if and when two components actually are equivalent (relative to a specific definition). Note that however dissimilar study and target, if some Hyp's components $\langle D, M, U\rangle$ and some Rep_k's components $\langle D_k, M_k, U_k\rangle$ were not understood to be comparable whatsoever, Rep_k would not be considered evidence for Hyp in the first place.

The threshold between sufficient similarity and unacceptable deviation is straightforwardly encoded as a component's membership of the respective equivalence class. Line 1 in (A*) can thus be understood as follows:

$$\begin{aligned}&\langle D, M, U\rangle \text{ is similar to } \langle D_k, M_k, U_k\rangle :\Leftrightarrow \\ &D_k \in [D]_\sim \text{ and } M_k \in [M]_\sim \text{ and } U_k \in [U]_\sim.\end{aligned} \tag{8}$$

Meaningfully evaluating the different components' membership of a given equivalence class calls for the distinction between similarity of *numeric* properties and similarity of *structural* properties. In the following I will outline possible paths one might take towards assessing similarity both of the numeric and the structural type.

2.3.1 Similarity of numeric properties

I will first consider the question of how to possibly compare arrays of numeric properties (or properties represented numerically, respectively). What it means to be a member of an equivalence class in that case can be expressed in a weighted geometric approach as the following criterion (with X schematically representing D, M, U):

$$X_k \in [X]_\sim :\Leftrightarrow \mathrm{d}(\overrightarrow{\beta} \circ \overrightarrow{x_k}, \overrightarrow{\beta} \circ \overrightarrow{x}) \leq \sigma_X. \tag{9}$$

Eq. 9 states that component X_k (e.g., some study population U_k) is similar to X (e.g., the target population U) for a purpose-oriented similarity evaluation $\sim$, explicated as distance between vectors of m properties $x_k^1, \ldots, x_k^m$ (of X_k) and $x^1, \ldots, x^m$ (of X)

[15] In the formal notation used here, $\sim$ denotes some (reflexive, transitive, and symmetric) equivalence relation (equivalence w.r.t. a given property) such that for a domain A, some object $a \in A$, and an equivalence relation $\sim$ on A: $[a]_\sim := \{x \in A \,|\, x \sim a\}$. The expressions $[D]_\sim$, $[M]_\sim$, and $[U]_\sim$ are to be understood as encoding each a specific equivalence relation, since – to be precise – each category comes with its own standards for how equivalence classes are to be generated. If standards are set high, e.g., D_k might be in the class $[D]_\sim$ only if it is identical with D, while comparing U and U_k will naturally demand flexibility for possibly very different populations. (I will not add a further index, though, to avoid notational clutter.)

above a certain distance threshold σ (relative to component X). To keep the criterion as generally applicable as possible, the elements of $\overrightarrow{x_k}$ and $\overrightarrow{x}$ should represent normalized values of the properties measured for comparison. When comparing two populations, for example, their relevant characteristics, such as, e.g., average age and population size, might be mapped onto scales of equal magnitude, to arrive at a balanced aggregate assessment. Additionally, the vectors to be compared are weighted by componentwise multiplication (denoted by '$\circ$') with relative weights $\beta^1, \ldots, \beta^m$, indicating the *relative importance* of each property for comparison, i.e., its contribution to the similarity measure.[16]

Of course, the distance measure d must be specified (Euclidean distance, Manhattan distance, etc.) and the set of properties to be compared chosen carefully – in particular, a simple distance measure might not be straightforwardly applicable if the selected properties are dependent. What counts as independent and what not is, again, ultimately to be decided in a purpose-oriented, context-sensitive way. In particular, different similarity standards could lead to different in/dependence requirements. A toy example may serve as an illustration of such dependencies: Consider the task of comparing different apples. If shape and color are to be compared, then a big red apple is more similar to a big green apple than to a small green apple. If, however, "naturalness" is used as similarity standard, then a big red apple might be considered more similar to a small green apple than to a big green apple, since the natural apple naturally ripens from small and green to big and red, and not to "unnatural" big and green. The latter case precisely illustrates the dependency between properties, and how such dependency is relative to the similarity standard to be employed.

Example.

The following simplified case equally shows advantages and shortcomings of such a geometric approach. Consider the assessment of a potential side-effect of a drug targeted at elderly people, who are likely to be using multiple medicines and likely to be affected by multiple diseases (note that those properties are dependent to a certain degree). Similarity between this specific target population and a test group (associated with report Rep_k) might be determined in comparison of property triples $\overrightarrow{u} = \langle u^1, u^2, u^3 \rangle$ and $\overrightarrow{u_k} = \langle u_k^1, u_k^2, u_k^3 \rangle$ where the properties to be compared are all numeric and partitioned into meaningful classes w.r.t. the current investigation. For example, it might be meaningful to distinguish the class of people who take 1 drug from the class of people who take 2 drugs, but it might not be particularly useful

[16] Componentwise multiplication of two vectors (also referred to as "Hadamard Product") multiplies vectors A and B (both of length n) element by element and returns a vector C (also of length n). Example: $\langle a, a, a \rangle \circ \langle 0, a, b, \rangle = \langle 0, 2a, ab \rangle$.

to draw the line between 11 and 12 multiple drugs. In our example, the selected characteristics $u_{(k)}$ of $U_{(k)}$ are partitioned as follows:[17]

$u^1_{(k)}$: (Average) age with integer values between 0 and 100 years of age;
$u^2_{(k)}$: (number of) multiple medicines in five classes:
0, 1, 2, 3-5, > 5 medicines;
$u^3_{(k)}$: (number of) multiple diseases in four classes:
0, 1, 2, > 2 diseases.

To be able to attach numbers to the vectors, we have to translate the description of U and U_k into concrete values step by step:

1. Suppose that the target population is described as 70 years old (on average), as taking 1 more medicine besides the tested drug (i.e., 2 in total), and as having more than 2 diseases considered relevant for the investigation (one targeted, the other possibly interacting).
2. The study is conducted on a population of 55 years on average with the same amount of medicines taken but with only one disease (the targeted one).
3. To be able to aggregate those properties on the same level, those characteristics are mapped onto the unit interval with respect to the partitions stated above. Moreover, for the current purpose, each class is simply represented by its central value (e.g., if the unit interval is partitioned into four intervals of equal length, the second interval, ranging from .25 to .5 is represented by its central value .375.).
4. Suppose further that – for the current investigation – average age is considered somewhat important (50%), whether multiple drugs are taken or not is considered highly important (100%), and the number of diseases is negligible (with importance of 0%).

The following tables show the step-wise calculation of the difference between weighted components needed for the determination of the distance between vectors $\overrightarrow{u}$ and $\overrightarrow{u_k}$:

	study population		
	class	normalized	quantized
u^1_k:	55		.55
u^2_k:	2		.5
u^3_k:	1		.375

	target population		
	class	normalized	quantized
u^1:	70		.7
u^2:	2		.5
u^3:	> 2		.625

	β	study	target	Δ
$u^1_{(k)}$:	.5	.275	.35	.075
$u^2_{(k)}$:	1	.5	.5	0
$u^3_{(k)}$:	0	0	0	0

[17] In this example, study and target are compared merely in terms of population characteristics $\overrightarrow{u_{(k)}}$. In general, though, differences between the substances and between the causal structures will also play a role in assessing the weight of a report, as explicated in Eq. 8. For sake of illustration it might be assumed here that substances and causal structures have been found equivalent w.r.t. the present purpose.

The upper tables show each property's class and its normalization in the unit interval (graphically with class partitions, and quantized using those classes' central values). The lower table shows the absolute differences between the individual components, weighted by corresponding β-values (relative importance). In our example, only *age* figures in the calculation of similarity between the two populations (since both groups take two medicines in total and the number of diseases is considered unimportant for the current comparison). If a Euclidean measure is to be used for calculating the distance, we arrive at the following value:

$$\mathrm{d}(\overrightarrow{\beta} \circ \overrightarrow{x_k}, \overrightarrow{\beta} \circ \overrightarrow{x}) = \sqrt{\left(\beta^1 \cdot (u^1 - u_k^1)\right)^2} = \sqrt{\left(.5 \cdot (.35 - .275)\right)^2} = .053 \qquad (10)$$

If, for comparison, the number of diseases were indeed to be taken into consideration with a relative importance weight of .25, the dissimilarity increases:

$$\begin{aligned} \mathrm{d}(\overrightarrow{\beta'} \circ \overrightarrow{x_k}, \overrightarrow{\beta'} \circ \overrightarrow{x}) = \\ \sqrt{\left(\beta^1 \cdot (u^1 - u_k^1)\right)^2 + \left(\beta^3 \cdot (u^3 - u_k^3)\right)^2} = \\ \sqrt{\left(.5 \cdot (.35 - .275)\right)^2 + \left(.25 \cdot (.625 - .375)\right)^2} = \\ \sqrt{.0375^2 + .0625^2} = .073 \end{aligned} \qquad (11)$$

A σ_U benchmark value for testing *sufficient similarity* with the target population (as formulated in Eq. 9) will have to be defined with respect to the theoretical maximum distance and to allowable deviations from the target characteristics. Once σ_U is fixed and *sufficient similarity* between study and target properties is shown, the report Rep_k can be flagged as relevant (with a high degree of confidence in α_k, i.e., 1 in the binary formulation), and (A*) will allow inference by analogy from relevant evidence to the hypothesis under consideration.

There are various ways to refine the distance measure and adjust it to one's needs. Two remarks on the difficult task of choosing a suitable measure shall conclude this section:

1. The example above illustrates that the chosen Euclidean distance measure completely disregards identical properties, even such with high relative importance, since it is defined to simply collect *deviations*. In the example, the identical properties u^2 and u_k^2 could be missing from the vectors $\overrightarrow{u}$ and $\overrightarrow{u_k}$ and could receive high or low relative importance – the Euclidean distance would return the same value in all these cases. This goes against the intuition "the more *identical properties* two property vectors share, the lower their distance". If this intuition is to be encoded formally, though, an additional weighting factor must be added to the distance measure. Each component x^i $(0 < i < m)$ of $\overrightarrow{x}$ might additionally be multiplied by its *relative* β*-contribution*, e.g., in the following way:

$$\mathrm{d}'(\overrightarrow{\beta} \circ \overrightarrow{x_k}, \overrightarrow{\beta} \circ \overrightarrow{x}) = \sqrt{\sum_{i=1}^{m} \left(\left(\beta^i \cdot (x_k^i - x^i)\right)^2 \cdot \frac{\beta^i}{\sum_{j=1}^{m} \beta^j} \right)}. \tag{12}$$

Using d' on our example case above returns a lower dissimilarity, thus reflecting the intuition that identical properties contribute to our similarity assessment (in decreasing dissimilarity). With $\sum_{j=1}^{m} \beta^j = 1.5$, we have:

$$\begin{aligned} \mathrm{d}'(\overrightarrow{\beta} \circ \overrightarrow{x_k}, \overrightarrow{\beta} \circ \overrightarrow{x}) = \\ \sqrt{\left(\beta^1 \cdot (u^1 - u_k^1)\right)^2 \cdot \frac{\beta^1}{1.5} + \left(\beta^2 \cdot (u^2 - u_k^2)\right)^2 \cdot \frac{\beta^2}{1.5}} = \\ \sqrt{\left(.5 \cdot (.35 - .275)\right)^2 \cdot \frac{.5}{1.5} + 0} = \\ \sqrt{.0375^2 \cdot \frac{1}{3}} = .022 \end{aligned} \tag{13}$$

And again, for comparison – for d' with relative importance of .25 for u_k^3, we get:

$$\begin{aligned} \mathrm{d}'(\overrightarrow{\beta'} \circ \overrightarrow{x_k}, \overrightarrow{\beta'} \circ \overrightarrow{x}) = \\ \sqrt{\left(.5 \cdot (.35 - .275)\right)^2 \cdot \frac{.5}{1.75} + \left(.25 \cdot (.625 - .375)\right)^2 \cdot \frac{.25}{1.75}} = .031 \end{aligned} \tag{14}$$

2. If, e.g., a certain property is required to be *identical* in study and target population in order to grant transferability of study results, partial identity, being central to the contrast account of similarity, can be expressed in a straightforward generalization of the geometric approach exemplified above: Instead of determining *sufficient similarity* for full vectors $\overrightarrow{u}$ and $\overrightarrow{u_k}$, those vectors are partitioned into sub-vectors. $\overrightarrow{u}$ and $\overrightarrow{u_k}$ are then considered *sufficiently similar* iff all distances of corresponding sub-vectors are lower than associated σ-values. For example, average age, number of medicines, and number of diseases (as in the case above) might be partitioned into sub-vector 1, average age plus number of diseases, and sub-vector 2 containing number of medicines only. If now the number of medicines taken is to be identical in target and study population (e.g., it might be required to be only the investigated drug), then the σ-value for sub-vector 2 is set to 0, while the σ-value for sub-vector 1 might be higher to allow for approximate correspondence.

2.3.2 Similarity of structural properties

When it comes to determining whether two structures (or, alternatively, sets of structural constraints) M_k and M are sufficiently similar, things are different: There is no straightforward way of encoding topology on a numeric scale. In particular, it depends very much on the given case what the comparison criteria are, and how they

possibly interact to jointly suggest a similarity ranking over a set of causal structures.

Nevertheless, the relevance of a given study for the investigated hypothesis crucially depends also on structural knowledge about the (hypothesized) causal associations in both study and target. For example, when an RCT on animals supports a dose-response relationship between drug D and adverse drug reaction E, this report can only be considered relevant if assuming a causal path from D to E is *compatible* with knowledge about the target's causal structure. And conversely, as soon as further studies show the association to be spurious in human drug users, the relevance of the animal study for the current investigation will be downgraded.

The following compilation presents a (non-exhaustive) list of structural properties to consider when comparing causal topologies. The respective questions can be understood as a heuristic survey for assessing structural similarity:

1. Shared **variables**: How many and which variables are shared between study and target? How influential are the ones that are not shared?
2. Causal **in/dependencies** of shared variables: Are the causal (conditional) in/dependencies of the variables shared by study and target in agreement? In other words, are the *causal ir/relevance relations* in study and target the same?[18] If not, is the disagreement resolvable (e.g., by including suspected confounders)?
3. Presence of **co-factors**: If a causal structure marks contributing causes or necessary pre-conditions – can these be identified both in study in target? Are the co-factors shared between study and target? How weighty are those that are not shared?
4. **Distance** between D and E: How many mediating variables lie on the path between D and E? If one structure contains more such mediators, does this reflect temporally or spatially higher actual distance? And do such mediators mark possible points for disruption by disablers, such that the causal process must be considered "less stable"?
5. **Interacting influences**: If E is influenced through multiple paths which jointly (but non-additively) produce E, then identifying these influences (and their interaction) in study and target is crucial for transferring predictions about the effect of interventions on D – this is particularly important in the case of masked *back-up mechanisms* present in the target.
6. **Differences** between study and target: Going beyond the points above, one might also utilize knowledge about how and where two causal structures differ. Information about differences at *key stages* along the causal path from D to E will be most helpful if those stages screen off influences of minor stages "upstream" along the causal flow (i.e., above or before). This idea is exploited in *comparative process tracing*, where two processes are investigated in stepwise comparison of their entities. Steel relates this method to external validity in Steel 2008, p. 89:

[18] See Pearl 2000, Sec. 7.3.3 for a discussion of causal relevance. Note, that I am distinguishing *causal relevance* as part of the causal knowledge (encoded as report nodes) from *epistemic* or *inferential relevance* (encoded as attributional *Rlv* weight nodes).

> [T]he greater the similarity of configuration and behavior of entities involved in the mechanism at [...] key stages, the stronger the basis for the extrapolation.
>
> The reliability of comparative process tracing depends on correctly identifying the points at which significant differences between the model and the target are likely to arise. Significant differences are those that would make a difference to whether the causal generalization to be extrapolated is true in the target.

The strategy summarized here is especially suitable for causal chains with such key variables – identifying precisely those variables (and describing deviating behavior in study and target) will help in deciding whether given study results provide an insight into the target. One of the virtues of comparative process tracing is indeed that it points to the portions of a causal structure that allow for an effective assessment of the similarity between study and target.

This list of structural properties is meant to provide a guide in making *implicit* assumptions about study and target *explicit*. Just like in the case of comparing numeric properties above, all these considerations feed into the purpose-oriented explication of a similarity measure which ultimately determines what *sufficient similarity* means in light of the investigated hypothesis.[19] As above, once *sufficient similarity* between structural properties of study and target is shown, a particular report Rep_k about M_k can be flagged as relevant (with a high degree of confidence in α_k, i.e., 1 in the binary formulation), and (A*) will allow inference *by analogy* from relevant evidence, summarized in Rep_k, to the hypothesized causal association, $D \copyright E$ in the causal structure M, thereby facilitating Bayesian confirmation:

$$P(D \copyright E \,|\, Rep_k, \alpha_k) \geq P(D \copyright E).$$

2.4 Extrapolating with good arguments and breaking the extrapolator's circle

I want to conclude Sec. 2 with a discussion of two points in the current debate on analogical reasoning in causal assessment. The first is Nancy Cartwright's worry about external validity, namely that even highly reliable RCT methodology only provides an insufficient warrant for the transferability of evidence from study to target (Cartwright, 2011; Cartwright and Stegenga, 2011). The second is a much-

[19] The way it is presented here, one's assessment of such similarity between study and target is obviously relative to the set of aspects included in one's considerations. There is an argument to be made here that possible further differences not considered should lower one's confidence in the similarity assessment. In principle, in the Bayesian framework employed, it is possible to add an unspecified counter-weight (much like an error term) in order to encode one's uncertainty about potentially neglected, though relevant differences between study and target. Yet, assigning a number to this weight is a subjective task again. Indeed, I would like to argue that such analogy-based arguments are inherently perspectival: They rest on a specific choice of relevant aspects (reasonably motivated) and a specific way of relating those aspects (non-arbitrarily). Thus, making the ingredients of such arguments explicit helps refining or potentially also refuting them.

discussed riddle about the soundness of analogical arguments – the 'extrapolator's circle' (Guala, 2010; Steel, 2008).

2.4.1 What can make RCT evidence relevant?

In her discussion of predictions about the effectiveness of policy interventions in the context of *evidence-based policy* (EBP), Cartwright points out that "[t]he EBP literature rates positive outcomes in well-conducted RCTs as gold standard evidence for effectiveness predictions" (Cartwright, 2011, p. 221). And she critically continues:

> Conventionally cited facts, like similarity between target and experimental situations, are then supposed to make this likely. But this is the wrong way to look at the relation between experimental results and the claims whose truth they bear on. Experimental results can help justify confidence that the same result – or that some different result – will hold elsewhere; i.e., they can be evidence for one of these claims. Whether they are evidence depends on whether they play the right kind of role in a good argument for that claim. Similarity, or just the right kind of dissimilarity, might play a role, but if so, only by fitting into a good argument.
>
> What then might a good argument look like that makes RCT results evidence for effectiveness predictions?

In this passage, Cartwright expresses precisely the worry that prompted Landes, Osimani, and Poellinger (Landes et al, 2017) to disentangle *reliability* and *relevance* as meta-evidential attributes. In order to address the question "What can make RCT evidence relevant?", Cartwright presents the following argumentative template (Cartwright 2011, p. 222; formatting adjusted):

> **A1** x plays a causal role in the principle that governs y's production there.
> **A2** x plays the same causal role here as there.
> **A3** The support factors necessary for x to operate are present for some individuals here.
> **Therefore:** x plays a causal role here and the support factors necessary for it to operate are present for some individuals.

In this argument, x is to be understood as a cause of y, 'here' indicates the target, 'there' the study. The four lines evidently resemble scheme (A) above: A2 and A3 establish similarity in the relevant (numeric, structural) aspects, and A1 encodes the RCT evidence (e.g., some effect size) to be transferred to the target in the conclusion. In the Bayesian framework presented above, the existence of such a *good argument* would be encoded as high *relevance*, consequently making the RCT result evidence for effectiveness predictions by boosting the weight of the respective evidential report.

2.4.2 The extrapolator's circle

In his discussion of mechanistic reasoning for the purpose of extrapolation, Daniel Steel (see Steel 2008, p. 78) presents the following challenge any viable account of extrapolation ought to address (see also Guala's comments in Guala 2010):

> [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.

In the account of Landes, Osimani, and Poellinger (Landes et al, 2017), this circle is broken since the proposed Bayesian framework can be used to probabilistically model *successive* evidence accumulation and amalgamation. At the beginning of the process, experimental results from basic science might contribute causal knowledge about a drug's metabolization – and even if this knowledge only illuminates part of the target mechanism, it can often be considered robust and highly relevant. Once animal studies come into play to answer more specific research questions, knowledge from previous unrelated studies (done on the same animals, maybe with similar substances) figures in forming a relevance estimate for the current investigation. Step by step, a picture of the target's causal structure emerges: The careful extrapolator can thus utilize the previous stage for his assessment of the next. While Steel proposes mechanism-based *comparative process tracing* (see also Sec. 2.3.2 above) as a solution to the extrapolator's circle, the Bayesian evidence-amalgamation approach presents a different type of 'process tracing' in providing a toolbox for tracing the dynamic process of evidence synthesis from a higher-level epistemological perspective.

The focus of this section lay on the confirmatory support of relevant pieces of evidence for a causal hypothesis ("upwards" in the evidence-amalgamating framework) – the next section will consider cases where an independently confirmed causal link serves for boosting one's confidence in a hypothesized relation.

3 Transferring knowledge from confirmed causal links

When Hill states that "with the effects of thalidomide and rubella before us we would surely be ready to accept slighter but similar evidence with another drug or another disease in pregnancy" (Hill, 1965), he refers to the case of an independently confirmed causal link providing support to a hypothesis still unsettled. Unlike in the case of learning from relevant evidence along the vertical upward arrow in Fig. 2, it is not an evidence report about an indicator of Hyp itself, but instead a well-tested second hypothesis Hyp^* that boosts belief in Hyp. Even though little might

be known about the mechanisms of substances D^* and/or D, Hyp is considered to share relevant *theoretical consequences* with Hyp^* (or variants thereof).

In order to formalize knowledge transfer across theoretical networks, our conceptual frame needs to be widened slightly. In a recent paper on analogical inference in physics, Dardashti et al (2017) discuss analogies between experimentally accessible test setups and potentially less accessible target systems we want to gain insights about. The authors introduce the formal concept of *analog simulation* for this purpose which shall be introduced briefly in the following before it is applied in the context of pharmacology.

Analog simulation bridges two basic frames: The source system is prepared, manipulated, and observed to make inferences about the target system. Let us introduce some terminology first to relate all concepts in a formal way:[20]

1. The target system T (a class of situations of interest) is to be modeled as $\mathscr{M}_T$ in a suitably chosen modeling framework $\mathscr{L}_T$;
2. $\mathscr{M}_T$ is constrained by certain background assumptions $\mathscr{A}_T$, summarizing theoretical and empirical knowledge as well as the domain of conditions $\mathscr{D}_T$ to which the model is intended to apply;
3. $\mathscr{M}_T$ can be used to predict phenomena P_T and will in turn be validated by evidence in accordance with P_T;
4. The accessible source system S is to be modeled as $\mathscr{M}_S$ in a suitably chosen modeling framework $\mathscr{L}_S$;
5. $\mathscr{M}_S$ is constrained by background assumptions $\mathscr{A}_S$, containing the domain of conditions $\mathscr{D}_S$ to which the model is intended to apply;
6. Just as on the T side, $\mathscr{M}_S$ can be used to predict phenomena P_S and will in turn be validated by evidence in accordance with P_S.

The source system S will now allow *analog simulation* of target T's behavior if (i) there exist exploitable structural similarities between $\mathscr{M}_S$ and $\mathscr{M}_T$ sufficient to define a syntactic isomorphism robust within the domains $\mathscr{D}_S$ and $\mathscr{D}_T$, respectively, and if (ii) this isomorphism is prompted by and based on a set of *model-external empirically grounded arguments*, abbreviated as MEEGA.

Figure 3 relates these elements in a conceptual graph: The rounded box on the left side contains all elements of the target frame, while the right box contains all elements of the source system. MEEGA prompt the establishment of a bridge between theoretical networks in the form of a syntactic isomorphism as translation between the systems' components.

In Dardashti et al (2017), the terminology is illustrated with an example from physics, where observations of phenomena P_S in table-top fluid systems boost confidence in theoretical assumptions $\mathscr{A}_T$ about gravitational phenomena described in

[20] In the following, I deviate from Dardashti et al (2017) in notational details.

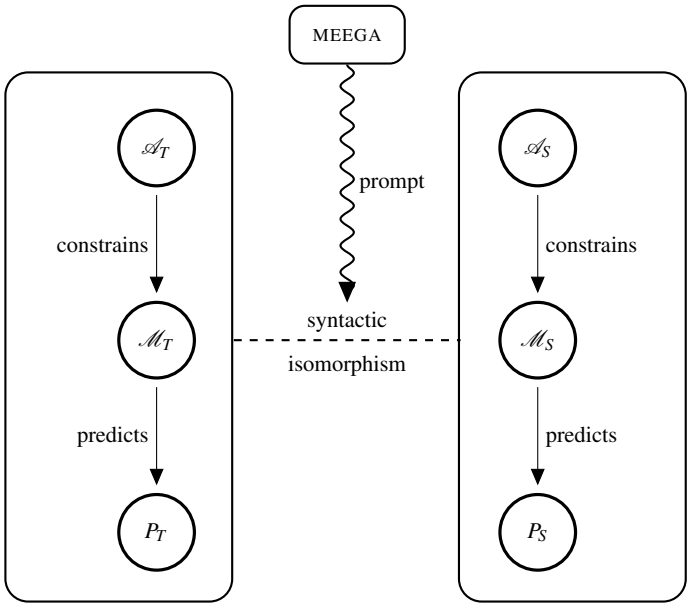

Fig. 3: The analog simulation scheme: Framework $\mathscr{L}_T$ (left box) is used to model target system T in model $\mathscr{M}_T$; source system S is accordingly treated in framework $\mathscr{L}_S$ (right box).

framework $\mathscr{L}_T$. The syntactic isomorphism (motivated by additional knowledge about the underlying physics of both frames) allows for the transfer of knowledge about acoustic Hawking radiation in the fluid system to Hawking radiation in black holes.

Now, the three-layered reconstruction of scientific domains on both sides of the syntactic bridge essentially represents the same conceptual categories as our layered reconstruction of inference in pharmacological research in Fig. 2: Scientific hypotheses entail system constraints which in turn predict (and are tested by) real-world phenomena. The syntactic isomorphism can be understood as a formal expression of similarity in terms of *partial identity under translation*. The set MEEGA can be understood to empirically, pragmatically, and semantically *prompt* the choice of relevant theoretical elements to be mapped onto each other. Syntactic isomorphism alone (unaware of semantic context) would be too weak a requirement for analogy since it can be used to translate far more models into each other than one would like to call *analog*. With the analog simulation scheme at hand, we can now trace the confirmatory boost of a well-tested causal link to an unsettled hypothesis in pharmacological research.

Fig. 4 shows the independently confirmed hypothesis Hyp^* on the right side, e.g., as in Hill's example, the confirmed link between thalidomide (D^*), also known

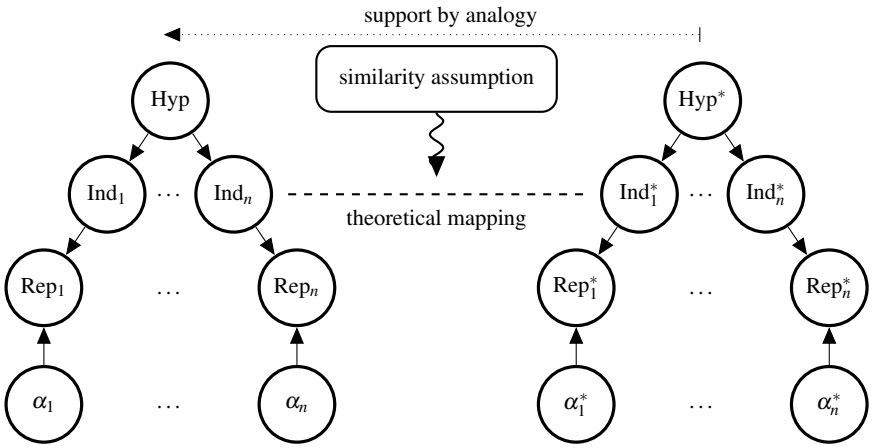

Fig. 4: Support by analogy from a second drug (i.e., from established knowledge about hypothesis Hyp^*) to the hypothesis under consideration.

as Contergan, and phocomelia (E^*). Frame-external similarity assumptions (e.g., about the population of pregnant women, or even new *evidence* for relevant relations shared by both M and M^*) suggest an understanding of theoretical variables Ind_k in terms of Ind_k^*, e.g., a characteristic dose-response curve hinting at functional properties shared by D and D^*. Having established such a mapping between relevant theoretical elements, updating one's beliefs about the variables Ind_k^* amounts to updating one's beliefs about the variables Ind_k and, moreover, to updating one's belief about Hyp (quite in the Bayesian sense).[21] Consequently, *old evidence* $\overrightarrow{Rep^*}$ about Hyp^* can indeed provide novel support for Hyp across theoretical networks, with

$$P(Hyp \,|\, \overrightarrow{Rep^*}) > P(Hyp). \tag{15}$$

How exactly this theoretical bridge between two frames might look and how a relevance filter is to filter out precisely those properties and those indicator variables relevant for the posited analogy, is a subject of current discussion in the philosophy of science; see, e.g., Dardashti et al (n.d) on analog simulation in Bayesian terms,

[21] As an illustration, consider the following: In many cases, *evidence for similarity* of the drug's causal effects comes from mechanistic knowledge, maybe in relating the molecular structure of the substances to known classes of biochemical processes. So, if D^* is known to be harmful because of its capacity to block some specific mechanism, and if this capacity is judged to be relevant in comparing D and D^*, then such blocking behavior should be part of Hyp's testable consequences Ind_k. Owing to differences in the investigated substances, the testable consequences of Hyp and Hyp^* are in general not *identical*, but can be related non-arbitrarily in motivating a specific *theoretical mapping*, i.e., some *isomorphism* at a suitably chosen level of description.

or Beebe and Poellinger (n.d.) on confirmation from analog models in formal extensions of Bayesian networks.

4 Confirmatory support from in silico simulation

When insights are to be gained about mechanistic workings of biochemical phenomena that are not directly observable in vivo, one strategy of choice is computational modeling and simulation. As in the cases of learning from relevant evidence (Sec. 2) and transferring knowledge from confirmed causal links (Sec. 3), using computational models to learn about biological, medical, or pharmacological facts is based on analogy between simulation and target system.

The argumentative strength of analog simulation in physics is based on the fact that two *physical* systems are related on the grounds of model-external background assumptions about common underlying *physical* principles. In the case of computational simulation, we seemingly loose two important aspects:

(M) Computational simulation does not link two physical systems but rather a physical and a *virtual* system; and
(N) the set of *model-external, empirically grounded assumptions* (MEEGA) is replaced by *model-internal, theoretical principles* in the implementation phase: The symbolic system $\mathscr{M}_S$ is constructed directly from $\mathscr{M}_T$'s background assumptions.

These aspects raise questions about virtues of materiality (M) and novelty of virtual outcomes (N). Both shall be addressed in the following.

Computer simulations follow their source code and will behave like programmed as virtual, artificial environments, lacking the material link (M).[22] The position of the skeptic about computer simulation is pointedly summarized by Diez Roux in her discussion of the distinction between observation-based and simulation-based causal inference in epidemiology (Diez Roux, 2015, p. 101):

> [...] there is a fundamental distinction between causal inference based on observations (as in traditional epidemiology) and causal inference based on simulation modeling. The traditional tools of epidemiology are used to extract (hopefully) reasonable conclusions from necessarily partial and incomplete (often messy) observations of the real world. [...] In contrast, when we use the tools of complex systems, we create a virtual world (based on prior knowledge or intuition) and then explore hypotheses about causes under the assumptions encoded in this virtual world. In the simulation model, we cannot directly determine

[22] This topic is a subject of current discussion in the philosophy of science with some authors regarding computational models as simply an implemented variant of scientific models as such (e.g., Frigg and Reiss 2009), while others emphasize as a special feature of computer simulations the possibility to experiment with such models as virtual test objects (e.g., Parker 2009 and Morrison 2015).

> whether X causes Y in the real world (because the world in which we are working is of our own creation); we can only explore the plausible implications of changing X on levels of Y under the conditions encoded in the model. In the real world, we have fact (what we observe) and we try to infer the counterfactual condition (what we would have observed if the treatment had been different). In the simulated world, everything is counterfactual in the sense that the world and all possible scenarios are artificially created by the scientist.

Nevertheless, even though computational models are virtual constructs, if they are to be employed for inference about our world they are required to be anchored in reality just as any classical scientific model. In a systematic review of successful agent-based computational models, Casini and Manzo (2016) aim to address Diez Roux's worries. In particular, they pin down various trends which have shown to increase the fruitfulness of such models. In the following I derive from their discussion three conditions that might even be understood as applicability criteria for causal inference from computational, agent-based models (ABMs) in general:

"Ideal" ABMs are to be

1. based on problem-related theoretical knowledge (rather than merely common-sense, mathematical, topological etc. assumptions),
2. shaped by data,
3. and iteratively calibrated by more data where the model shows weaknesses.

Fig. 5 illustrates the relation between target system and computational model in the analog simulation scheme, which I modify here to accommodate the anchoring concept: The source system (our computational model) inherits all theoretical background assumptions from the target side. The arrow from $\mathscr{A}_T$ directly to $\mathscr{M}_S$ represents the preparation of the simulation system as logical cross-dependency, which can be called *model-internal* in a sense: Without *model-external* motivation provided by MEEGA, the rigid link from $\mathscr{A}_T$ to $\mathscr{M}_S$ expresses the instantiation of the theoretical assumptions $\mathscr{A}_T$ in a concrete (computational) model $\mathscr{M}_S$ *within* the constraints of $\mathscr{A}_T$. Whenever incoming data alters the background assumptions $\mathscr{A}_T$, $\mathscr{M}_S$ (the implementation) can (must) be revised accordingly.

What this picture reveals, though, is that – as formulated above in (N) – the *prediction* of P_S now turns into plain *inference* from $\mathscr{A}_T$. Consequently, it would not make sense to use P_S for the confirmation of hypothesis $\mathscr{A}_T$. Whatever theory of confirmation is to be employed, analog reasoning will only be justified if source and target systems are kept sufficiently independent: The computational model must not be determined by the target system in order to leave room for the possibility of *disconfirmation*.[23]

What are our options? – The direct, logical dependence between $\mathscr{A}_T$ and $\mathscr{M}_S$ may be disrupted by a reintroduced $\mathscr{A}_S$ in different ways: (i) $\mathscr{A}_S$ might be built

[23] In the context of modeling with Bayesian networks, this demand is captured in the requirement that all variables represent distinct events.

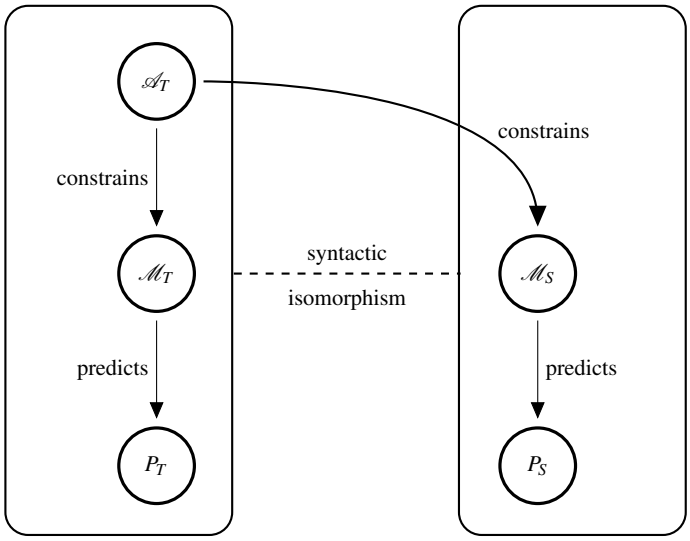

Fig. 5: The virtual source system inherits the target's theoretical background assumptions; the arrow from $\mathscr{A}_T$ to $\mathscr{M}_S$ represents the preparation of $\mathscr{M}_S$ as logical cross-dependency.

upon a lower-level/finer-grained theory than $\mathscr{A}_T$, (ii) $\mathscr{A}_S$ might draw on a different set of parameters, (iii) $\mathscr{A}_S$ might incorporate theories from different domains, (iv) $\mathscr{A}_S$ might utilize databases generated from previously conducted material simulations. Of course, the list is certainly not exhaustive, and strategies might possibly be combined. Once the cross-link from $\mathscr{A}_T$ to $\mathscr{M}_S$ is disrupted by integrating external assumptions, P_S regains its confirmatory support towards $\mathscr{A}_T$.[24] In a case study on computational modeling of cell proliferation mechanisms in systems biology, Osimani and Poellinger (n.d.) identify different ways in which a computer simulation can provide novel insights about the object of interest and confirmatory support to a scientific hypothesis:

1. The bottom-up combination of independently secured pieces of knowledge may produce unexpected results precisely because the components' interaction might influence the functioning of the whole mechanism.
2. Combining mechanistic knowledge from different sources might reveal surprising insights about hidden mechanisms in mismatches between virtual measurements and expectation.[25]

[24] This strategy introduces a secondary set of *model-external, empirically grounded arguments* in the picture which is first motivated by the logical cross-link between the two frames and later guided by anchoring considerations. See Osimani and Poellinger (n.d.) for a detailed reconstruction of model creation, verification, and validation for computer simulation in systems biology.

[25] For a discussion of surprise in computer simulation see Parke (2014).

3. Moreover, novel insights might be generated when phenomena (potentially not predictable from a small set of basic rules) emerge in iterations of a complex computer simulation.

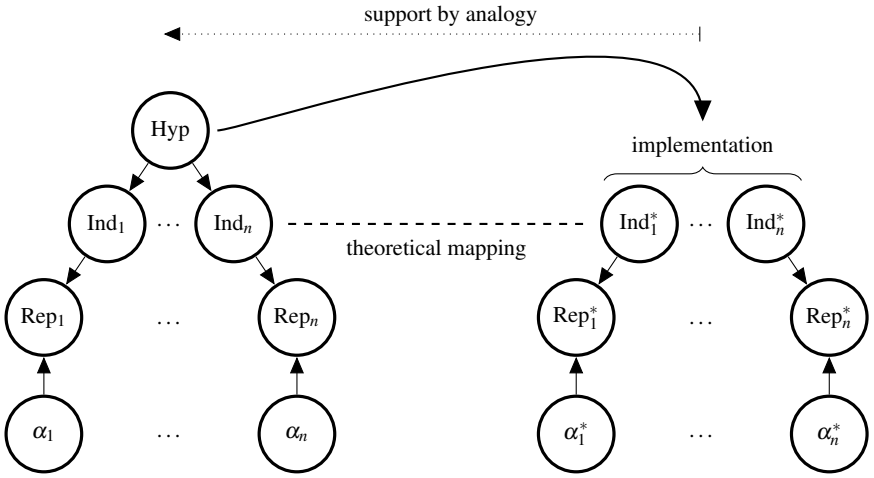

Fig. 6: Support by analogy from a computational model of the target system to the hypothesis about the target.

Fig. 6 illustrates the conceptual dependencies in confirmation by analogy in our layered reconstruction of pharmacological research: Hyp is implemented by using the theoretical background assumptions about $D \copyright E$ in the first version of the computational model. By refining and enriching this model during verification and validation, the logical cross-dependency between Hyp and its implementation is disrupted to an extent that allows for novel 'observations' during runs of the simulation, such that analogy can be used for the confirmation of Hyp once more.

5 Conclusions

The evidence-amalgamation framework, introduced in Sec. 1, opens the possibility of tracing the dynamics of analogical reasoning across distinct epistemological categories: theoretical hypothesis (Hyp), testable indicators (Ind), and evidence reports (Rep). Within this layered reconstruction of scientific inference, the concept of Bayesian confirmation can be utilized to formulate precisely how a causal hypothesis about a drug's potentially harmful side-effects is confirmed or disconfirmed.

This paper distinguishes three analogy-based inference patterns significant in pharmacological research:

1. **Inference from relevant reports:** When the conditions of a given study correspond to the intended application of the investigated hypothesis, reports about the study are marked as relevant for the hypothesis, thus facilitating knowledge transfer from evidence to hypothesis. In Sec. 2, study and target conditions are broken into three components: the drug itself (D), the causal model implicit in the hypothesis (M), and the respective population (U). Pair-wise comparison is based on a similarity measure to be chosen w.r.t. the nature of the investigation; e.g., the components might be compared using a geometric measure of similarity as given by the distance between property vectors like in the example case. Obviously, some important decisions are to be taken outside the framework presented, though: Questions as to how to arrive at selection criteria for the properties to be compared are left for another paper.
2. **Inference from established causal knowledge:** Sec. 3 discusses analogical inference from a second well-tested hypothesis Hyp^*. In this case, the connection between source and target is not established via similarity but across a syntactic isomorphism between the hypotheses' consequence sets, i.e., a theoretical mapping on the indicator level. Once this bridge is defined (motivated and justified by model-external empirically grounded arguments), evidence for Hyp^* (the established hypothesis) will also boost confidence in Hyp (the hypothesis under investigation).
3. **Inference from computational models:** If the analog system is a virtual, computational model of the investigated hypothesis, the bridge between source and target is not motivated by model-external considerations but much rather by model-internal constraints. Sec. 4 discusses how an *artificial* system can possibly provide support for a causal hypothesis about an *actual* drug with *real* risk.

While the paper focuses on hypothesis testing for the purpose of risk assessment in pharmacology, the second and third pattern in the list above make the role of analogical reasoning in the formulation of a hypothesis obvious. In particular, a computer simulation will support an investigated hypothesis if the codebase is not merely constructed from theoretical assumptions about the target hypothesis, but infused with further theoretical considerations or additional sources of knowledge, thereby breaking the logical dependency between source and target (see Sec. 4). Once a theoretical bridge is established on the indicator level, unexpected, surprising, possibly unpredictable evidence reports about an analog system will propel hypothesis discovery and theory revision. Indeed, this extends beyond computational modeling: When Otto Schaumann created meperidine, the first fully synthetic opioid pain medication, in 1937, he observed that meperidine and morphine produced similar physiological signs when administered to rats in lab experiments. In addition, meperidine was known to share chemical structural properties with morphine. Schaumann consequently (and rightly) hypothesized that meperidine also shares morphine's narcotic effects (see Bartha 2013). This further episode of successful inference by analogy illustrates the peculiar nature of pharmacology, integrating

different levels of description from chemical structure to clinical observation. Justifying analogical arguments by reconstruction calls for a framework capable of accommodating heterogeneous sources of evidence that allows tracing confirmatory support across distinct epistemological categories – possibly also from analog systems. The collection of modules presented in this paper may serve as a toolbox for such justification in the scientific dialog between hypothesis testing and theory revision.

Acknowledgements This paper was presented at workshops and conferences in Munich, Sydney, Groningen, Bologna, Bochum, and Exeter. I greatly benefited from the comments and suggestions made by the audiences, and I am particularly thankful for personal discussions with Cameron Beebe, Lorenzo Casini, Radin Dardashti, Stephan Hartmann, Adam La Caze, Jürgen Landes, Barbara Osimani, Jan-Willem Romeijn, Karim Thébault, Naftali Weinberger, and Michael Wilde, whose valuable comments helped me clarify my aims and shape the final version of this paper.

References

Bartha P (2010) By Parallel Reasoning: The Construction and Evaluation of Analogical Arguments. Oxford University Press

Bartha P (2013) Analogy and analogical reasoning. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy, fall 2013 edn

Beebe C, Poellinger R (n.d.) Bayesian Confirmation from Analog Models, forthcoming.

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

Britton OBOA, Van Ammel K, Lu H, Towart R, Gallacher D, Rodriguez B (2013) Exeperimentally calibrated population of models predicts and explains intersubject variability in cardiac cellular electrophysiology. Proceedings of the National Academy of Sciences of the United States of America (110):E2098–105

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafo MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience 14:365–376, URL http://dx.doi.org/10.1038/nrn3475

Cartwright N (2011) Predicting what will happen when we act. what counts for warrant? Preventive Medicine 53(4):221 – 224, DOI https://doi.org/10.1016/j.ypmed.2011.08.011, URL http://www.sciencedirect.com/science/article/pii/S0091743511003008, special Section: Epidemiology, Risk, and Causation

Cartwright N, Stegenga J (2011) A Theory of Evidence for Evidence-Based Policy. In: Dawid P, Twinning M William Vasilaki (eds) Evidence, Inference and Enquiry, OUP, chap 11, pp 291–322

Carusi A, Burrage K, Rodriguez B (2012) Bridging experiments, models and simulations : an integrative approach to validation in computational cardiac electrophysiology. American Journal of Physiology - Heart and Circulatory Physi-

ology 303(2):H144–H155, DOI 10.1152/ajpheart.01151.2011, URL http://eprints.qut.edu.au/75632/

Casini L, Manzo G (2016) Agent-based models and causality: A methodological appraisal. The IAS Working Paper Series (Linköping University Electronic Press) (7), URL http://liu.diva-portal.org/smash/get/diva2:1058813/FULLTEXT01.pdf

Chan AW, Altman DG (2005) Epidemiology and reporting of randomised trials published in PubMed journals. The Lancet 365(9465):1159–1162, URL http://dx.doi.org/10.1016/S0140-6736(05)71879-1

Dardashti R, Thébault K, Winsberg E (2017) Confirmation via analogue simulation: What dumb holes could tell us about gravity. The British Journal for the Philosophy of Science 68(1):55, DOI 10.1093/bjps/axv010, URL http://dx.doi.org/10.1093/bjps/axv010, /oup/backfile/content_public/journal/bjps/68/1/10.1093_bjps_axv010/2/axv010.pdf

Dardashti R, Hartmann S, Thébault K, Winsberg E (n.d) Confirmation via analogue simulation: A bayesian analysis, forthcoming.

Diez Roux AV (2015) The virtual epidemiologist – promise and peril. American Journal of Epidemiology 181(2):100–102

Doll R, Peto R (1980) Randomised controlled trials and retrospective controls. British Medical Journal 280:44, URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1600504/

Frigg R, Reiss J (2009) The philosophy of simulation: hot new issues or same old stew? Synthese 169(3):593–613

Geiger D, Verma T, Pearl J (1990) Identifying independence in bayesian networks. Networks 20(5):507–534

Guala F (2010) Extrapolation, analogy, and comparative process tracing. Philosophy of Science 77(5):1070–1082

Hesse MB (1952) Operational Definition and Analogy in Physical Theories. British Journal for the Philosophy of Science 2(8):281–294, URL http://www.jstor.org/stable/686017

Hill AB (1965) The Environment and Disease: Association or Causation? Proceedings of the Royal Society of Medicine 58(5):295–300

LaFollette H, Shanks N (1995) Two Models of Models in Biomedical Research. Philosophical Quarterly 45(179):141–160, URL http://www.jstor.org/stable/2220412

Landes J, Osimani B, Poellinger R (2017) Epistemology of Causal Inference in Pharmacology. Towards an Epistemological Framework for the Assessment of Harms. European Journal for Philosophy of Science DOI 10.1007/s13194-017-0169-1, URL http://dx.doi.org/10.1007/s13194-017-0169-1

Lewis D (1973a) Causation 70(17):556–567

Lewis D (1973b) Counterfactuals, 2nd edn. Wiley-Blackwell

Luján JL, Todt O, Bengoetxea JB (2016) Mechanistic Information as Evidence in Decision-Oriented Science. Journal for General Philosophy

of Science 47(2):293–306, URL http://dx.doi.org/10.1007/s10838-015-9306-8

Morrison M (2015) Reconstructing Reality: Models, Mathematics, and Simulations. Oxford University Press

Osimani B, Poellinger R (n.d.) A Protocol for Model Validation and Causal Inference from Computer Simulation, forthcoming.

Parke EC (2014) Experiments, simulations, and epistemic privilege. Philosophy of Science 81(4):516–536

Parker WS (2009) Does matter really matter? computer simulations, experiments, and materiality. Synthese 169(3):483–496

Paul LA, Healy K (2016) Transformative treatments. Noûs 50(4)

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

Poellinger R (n.d.) On the Ramifications of Theory Choice in Causal Assessment: Indicators of Causation and Their Conceptual Relationships, forthcoming.

Revicki DA, Frank L (1999) Pharmacoeconomic Evaluation in the Real World. PharmacoEconomics 15(5):423–434, URL http://dx.doi.org/10.2165/00019053-199915050-00001

Shepard RN (1980) Multidimensional scaling, tree-fitting, and clustering. Science 210(4468):390–398, DOI 10.1126/science.210.4468.390, URL http://science.sciencemag.org/content/210/4468/390, http://science.sciencemag.org/content/210/4468/390.full.pdf

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

Tversky A (1977) Features of similarity. Psychological Review 84(4):327–352

Unruh WG (2008) Dumb holes: analogues for black holes. Philosophical Transactions of The Royal Society A 366:2905–2913, URL http://dx.doi.org/10.1098/rsta.2008.0062

Upshur R (1995) Looking for Rules in a World of Exceptions: reflections on evidence-based practice. Perspectives in Biology and Medicine 48(4):477–489, URL http://dx.doi.org/10.1353/pbm.2005.0098

Weisberg M (2012) Getting Serious about Similarity. Philosophy of Science 79(5):785–794

Weisberg M (2013) Simulation and Similarity: Using Models to Understand the World. Oxford University Press

Worrall J (2007) Evidence in Medicine and Evidence-Based Medicine. Philosophy Compass 2(6):981–1022, URL http://dx.doi.org/10.1111/j.1747-9991.2007.00106.x