Brit. J. Phil. Sci. **56** (2005), 791–807

Scientific Coherence and the Fusion of Experimental Results

David Danks

ABSTRACT

A pervasive feature of the sciences, particularly the applied sciences, is an experimental focus on a few (often only one) possible causal connections. At the same time, scientists often advance and apply relatively broad models that incorporate many different causal mechanisms. We are naturally led to ask whether there are normative rules for integrating multiple local experimental conclusions into models covering many additional variables. In this paper, we provide a positive answer to this question by developing several inference rules that use local causal models to place constraints on the integrated model, given quite general assumptions. We also demonstrate the practical value of these rules by applying them to a case study from ecology.

1 Experimental scope in applied sciences

Total photosynthetic material has increased globally in recent years (though with local decreases), and one might naturally wonder why. In a recent paper in *Science*, Nemani et al. ([2003]) focused on some of the potential causes of global vegetation growth during the past 20 years. Their analysis focused on only four variables: growing season average temperature, vapor pressure deficit, solar radiation, and net primary production (photosynthetic material). Their study considered only a limited variable set because of (a) the global scale of their analysis, and (b) the relatively long study period (18 years). Despite this limited scope (in terms of variables), their study gives substantial support to the hypothesis that the first three variables are causes of the last, and helps to clarify the functional form of those dependencies. At the same time, they explicitly note that there are many causally relevant variables that were ignored in their study, such as vegetation

doi:10.1093/bjps/axi146
Advance Access published on October 31, 2005.

decomposition rates and land-use changes during the past two decades. In this case, other studies will be necessary to determine the influence of these other variables; that is, the process of discovering the various causes of vegetation growth is fragmented.

This kind of fragmentation—studying only a few variables at a time, even though they may interact with other variables addressed in separate studies—is typical of the applied sciences, though for different reasons in different domains and experiments (e.g. technical versus financial versus ethical). In the climate sciences, some experiments/observational studies specialize in global heat and temperature, some in atmospheric phenomena, some in land phenomena, some in ocean phenomena, some in general circulation models, and so on. The social sciences are no different. Econometric models and studies of the United States economy share some variables, but not others, with econometric models of the United Kingdom. Datasets collected about poverty, employment, and almost any other social issue, as well as subsequent analyses of that data, typically share some variables but not others. Medical data collections, and studies based on them, exhibit similar fragmentation.

This patchwork character of applied science threatens its conclusions. In most of the examples just mentioned, we cannot perform experiments in which the hypothetical causes are deliberately randomized or otherwise manipulated, and so must generate or estimate models[1] based on both our observations and background beliefs. But, of course, correlation is not causation. If potentially relevant variables are knowingly omitted from a study, then we seemingly cannot be confident that the associations found among the variables that *are* considered are not due, wholly or in part, to omitted variables. We are forced to ask: how, if at all, can such fragments—these patches of perhaps tentative scientific results—be joined together into coherent models in which we can have some confidence? How, if at all, can multiple instances of local learning (i.e. an experiment on a small set of variables) be fused together into a global theory (i.e. a causal structure/set of interactions on a much larger set of variables)?

For a possible response, we can look towards work over the past 20 years by other philosophers of science, computer scientists, and statisticians to develop a framework—causal Bayesian networks—for representing causal hypotheses and extracting them from observational data and background knowledge. This framework and the associated search procedures have slowly

[1] No particular theoretical baggage is implied by the use of the word ‘model’. We will relatively interchangeably use the terms ‘model’, ‘theory’, and ‘structure’. Nothing we say will depend on the distinctions that are often drawn among these terms.

gained increasing acceptance in statistics, and have produced results in many sciences—both applied and 'pure'.[2] One might hope that this framework would enable us to provide an informative, normative way of stitching together the fragments of scientific knowledge in various applied sciences.[3] Most of the remainder of this paper considers the following three, more specific, questions:

1. Given a set of local, partially overlapping causal models, what constraints—if any—can be placed on an integrated model of all of the variables?
2. How can the introduction of background knowledge aid the construction of this integrated structure?
3. Given a set of local, partially overlapping causal models, what series of experiments will most efficiently determine the full, integrated causal model (while still respecting limits on the types of experiments that we can perform)?

Note that we restrict ourselves here to the normative dimension of this problem, and leave aside the obvious descriptive analogues of these three questions. We focus here on how applied scientists *could* stitch together the fragments, not whether they actually *do* stitch in this manner. Also, there are important interconnections among these questions: the second and third questions are only interestingly novel if there is a positive answer to the first question. If nothing can be learned about the integrated structure from the local models, then the latter two questions reduce to the well-studied problems of background knowledge incorporation and experiment planning in a state of ignorance about the integrated model. Perhaps surprisingly, it will turn out that we can provide partial positive, non-trivial answers to all three of these questions.

Bovens and Hartmann ([2002]) focused on a related problem: how can we fuse or unify the results of repeated measurements of one or more predictions

[2] Some examples of successful application of the causal Bayes net framework: biology (Shipley [2000]), economics (Bessler [2003]), educational research (Conati et al. [1997]), cognitive psychology (Waldmann and Martignon [1998]; Steyvers et al. [2003]; Danks, Griffiths and Tenenbaum [2003]), developmental psychology (Gopnik et al. [2004]), genetics (Smith et al. [2002]; Danks et al. [2003]), mechanical engineering (Lerner et al. [2002]), medicine (Cooper et al. [2000]), metaphysical accounts of causation (Woodward [2003]), mineral identification (Ramsey et al. [2002]), neuroscience (Glymour [2002]), and space physics (Waldemark and Norqvist [1999]).

[3] The problem of inferring causal relations from statistical evidence—particularly within the framework of causal Bayes nets—has been much discussed in recent philosophical literature (see, e.g. McKim and Turner [1997]; Cartwright [1999]; Glymour [1999]). This paper will sidestep that debate. Rather than focusing on the question of 'what do we need to assume to infer causation?', we will attempt to answer the question of 'what do we do with (local) causal models/structures once we have them?'

of a particular scientific hypothesis? That is, they provide an account of hypothesis confirmation from multiple experiments (or repetitions of the same experiment), under the assumption that our instruments are less than completely reliable (even if only because of inherent noisiness). Their work also used the framework of Bayesian networks, and can be viewed as a necessary precursor to the results in this paper. The three questions above all take for granted that some, perhaps quite minimal, causal information can be learned from experimental results. Bovens and Hartmann's ([2002]) work—among many interesting results—demonstrates how and when such learning is possible with noisy, possibly quite unreliable, measurement instruments.

In the next section of the paper, I provide (qualitative) inference rules for fusing the various local causal models, and briefly discuss how those rules provide partial answers to the second and third questions above. I then show how these rules can be applied to a simplified, but concrete, example in Section 3. The final section applies the inference rules to a particular case study from ecology, and draws some broader conclusions about the proposed partial solution to this problem of generating coherent global causal structures from disparate local causal models.

2 Fusing the results of experiments[4]

A completely standard rule of scientific inference is: 'If we experimentally change the state of *X* and *Y* does not change, then *X* does not cause *Y*'. There are, of course, situations in which this rule might fail (e.g. Hesslow's [1976] example of birth control pills and thrombosis; see also Cartwright's [1989] discussion of this case), but we can give a relatively precise characterization of the situations in which the rule will fail; determining whether this rule will yield correct information in a particular situation is itself a testable scientific hypothesis. So, for example, if we want to know whether running causes weight gain/loss, we could randomly assign people to either run three miles a day or minimize their physical exertion. If the average changes in weight in the two groups are not significantly different, then we conclude that running unfortunately does not cause weight change (unless one or more of the usual assumptions are violated, such as the uniformity of the two populations, the absence of exactly offsetting causal paths, etc.).

More typically, however, we are not interested simply in causation, but more specifically *direct* causation (relative to a particular system of variables).

[4] Although this section is written relatively informally, all of the results can be expressed precisely in terms of causal Bayes nets (see Danks [2002], [2003] for the exact details).

Intuitively, *X* directly causes *Y* relative to a set of variables when *X* exerts a causal influence on *Y* that does not pass through any of the other variables in the system. The previous inference rule draws no distinction between indirect and direct causation. The more commonly used inference rule—which does separate direct and indirect causation—is roughly: 'If we experimentally change *X*'s state, hold fixed (in some way) the other variables in the system, and *Y* does not change, then *X* does not directly cause *Y*'. The exact nature of the 'holding fixed' might vary across experiments: sometimes experimental clamping can be performed, other times we match individuals between the populations. And once again, we can quite precisely state the general assumptions that must hold for this inference rule to be asymptotically correct (or more precisely, different versions of this rule for different methods of 'holding fixed'). To continue our earlier example, we might (impractically) try to control experimentally our subjects' diets by forcing everyone in the study to eat exactly the same food. Or we might ensure that every subject in the running condition had a counterpart of the same weight in the sedentary condition. If the weight changes are not significantly different in either of these two cases, then we conclude that running is not a direct cause of weight change relative to the studied system (running, weight change, and diet or initial weight, respectively).

We can further extend the above inference principles from experimental to observational data. That is, we can draw (partial) conclusions about the absence of causal influence based on particular patterns of associations and independencies. In particular, we (tentatively) conclude that *X* does not directly cause *Y* when *X* and *Y* are independent conditional on (some subset of) the other variables in the system. The existence of observational inference rules is particularly important for many of the applied sciences, since experimental manipulations are often impossible, whether for financial, technical, or ethical reasons. The general assumptions required for the (asymptotic) correctness of these observational inferences are stronger than those required earlier, but they are again testable scientific hypotheses. So, in our toy example, we could measure, e.g. the running habits, recent weight change, and diet of many individuals in the population. Given some assumptions, if running and weight change are independent conditional on every measured value of diet, then we would conclude that running does not (directly) cause weight change.

Of course, all of these inference rules are instances of 'local' learning, as understood in the previous section: they all apply to cases in which we have data over all of the variables being studied, and there is no integration to be performed. But there is a thread running through these rules that we can exploit when stitching together local models: absence of association (conditional on some set) implies absence of (direct) causation (given some general, but testable, assumptions about the system).[5] And notice that the conditional independence of two variables does not change simply because we happen to measure more variables. The independence might change if we include those additional variables in the conditioning set, but independence conditional on *some*—not all—variables is all the various inference principles require to conclude absence of causation. Thus, when we learn about absence of causation in a local model, that absence will translate to the integrated global model. We can state this inference rule for stitching together local models more precisely as:

> **Inference rule 1:** A local conclusion that *X* does not directly cause *Y* relative to some system **S** holds relative to any supersystem **S*** (i.e. a system whose variables are a superset of the variables in **S**).

To finish the above example using this inference rule, if we (locally) conclude that running does not directly cause weight change relative to some system (e.g. running, weight change, and diet), then this conclusion holds for all super-systems of that one (e.g. if we measure metabolic rate in addition to these other three variables). As a side note, it is important to emphasize that 'inclusion of variables in a set' refers only to which variables are measured, and not which are assumed to be present (or absent), or which are experimentally manipulated, or even which (if any) must be included in the conditioning set for any independence test.[6]

For a more realistic example of this principle's application, suppose counterfactually that Nemani et al. ([2003]) had concluded that one of their variables (e.g. solar radiation) was not a (direct) cause of net primary production because those two variables were (perhaps conditionally) uncorrelated during the study period. In this case, the above inference rule would license us to conclude that (given certain testable assumptions) solar radiation does not directly cause net primary production in any model with additional variables (e.g. one that incorporated the influence of wind, or recent changes in land

[5] Technically minded readers might recognize this 'common thread' as essentially a colloquial statement of statement of Spirtes *et al.'s* ([2000]). Faithfulness assumption (called Stability in Pearl [2000]). However, since Faithfulness is actually one of the 'precise general assumptions' alluded to earlier, this might all appear disingenuous, since we seemingly 'derive' *X* from the assumption that *X*. However, the initial discussion in this section should not be read as 'proving' that the common thread is true, but rather as trying to argue that it is ubiquitous in scientific inference.

[6] That being said, 'inclusion of variables in a set' *does* include reference to the observed values of those variables. In particular, the statistical judgment that two variables are independent holds only for the observed values of those variables. This condition becomes particularly important when we have skewed datasets in which one or more variables cover only a small part of their range (e.g. income and education values among people on welfare). The inference of lack of causation between *X* and *Y* might not hold for a supersystem if *X* and *Y* also take on previously unobserved values in the supersystem (e.g. income, education, and housing values—the supersystem—in the entire U.S., in which they will take on a larger range of values). Thanks to an anonymous reviewer for emphasizing this point.

use patterns). And so, as we try to stitch together the many ecological models of vegetation growth that emerge from different experiments, we would have a constraint on the possible (causal structure of the) full model.

This inference rule is interesting, since it shows that the problem of stitching together local knowledge is not hopeless: some local information is reusable, even when we consider a larger model. However, this rule is—in a sense—itself highly local, since it uses information about the absence of causal connections from only one particular experiment. By assumption, though, we have multiple local models, and so we are not restricted to just this information. And we can place further constraints on the integrated model when we simultaneously consider these multiple models. The inference rule for these further constraints is best introduced through an illustrative example.

Suppose (local) experiments have yielded the following two causal structures: $X \rightarrow A \leftarrow Y$, and $A \rightarrow C \leftarrow B$, where '→' means direct causation (relative to the system of variables). That is, X and Y jointly cause A, but are independent of each other (so there is no causal connection between them by the earlier principles), and A and B jointly cause C, though they are also independent of each other. Now consider what constraints can be placed on the integrated model for all five variables. Specifically, consider variables X and B. There are four possible causal connections between B and X: (i) B causes X; (ii) X causes B; (iii) there is an unobserved common cause of X and B; or (iv) there is no causal connection between them at all. Note that these are not all mutually exclusive possibilities: conceivably, both X causes B and there is an unobserved common cause of the two variables. Now consider each of the first three possibilities. If B is a cause of X in the integrated model, then B is a cause of A (though indirectly through a variable, X, that was not measured in the second experiment). If B is a cause of A, then B and A should be unconditionally associated. Alternately, if X is a cause of B or there is an unobserved (in either experiment) common cause of X and B, then B and A have an unobserved common cause in the second experiment, which will induce an association between them (just as, e.g. barometer readings and storms are correlated when we do not observe the air pressure). However, since A and B are known to be unconditionally independent (from the second experiment), none of these three possibilities can be actual. Thus, the only remaining alternative is no direct causal connection at all between X and B. An analogous argument holds for Y and B. Thus, by simultaneously considering both local models, we can place two further constraints on the (causal structure of the) full model: there is no direct causal connection between X and B, or between Y and B.

This result is, in some ways, quite astonishing. We have eliminated the possibility of a causal connection between two variables (X and B, or Y and B) without those two variables ever appearing in the same dataset or experiment. That is, we have *no* datapoints (experimental or observational) with measurements for both X and B, or both Y and B, but we can still determine that there cannot be a causal connection between them. One might naturally have thought a priori that nothing could be learned about possible causal connections between variables that are never jointly measured; that intuition turns out to be incorrect (given certain testable assumptions about the data and the generating process).[7]

This is an intriguing example, but not sufficient by itself to establish a second general inference rule. To give the more general principle, we first need a semi-technical definition. We define 'X is a definite cause of Y' iff there is some (possibly empty) chain of intermediate variables $Z_1, \ldots, Z_n$ such that X is a cause of Z_1, Z_1 is a cause of $Z_2, \ldots,$ and Z_n is a cause of Y. Note that X, Y, and the Z's need not all appear in the same local model, but that this definition assumes the transitivity of causal relations. The question of the transitivity of causation has been much-discussed in the philosophical literature (e.g. Hitchcock [2001]), but I wish to sidestep those debates here. Instead, I simply note that the successful application of this rule depends upon transitivity. If transitivity always holds, then this rule always works (given the other assumptions); if transitivity sometimes fails, then the application of this rule will depend on case-by-case judgments, and the general usefulness of the rule will depend on the empirical frequency of non-transitive causal relations. Given this definition, we can give the following inference rule:

> **Inference rule 2:** Suppose there is no causal connection between X and Y in some local model because X and Y are unconditionally independent. Then there are no causal connections in the integrated model between X and the definite causes of Y, or between Y and the definite causes of X.

This inference rule simply formalizes the arguments used for the previous example: if there were some causal connection between X and a definite cause of Y, then X and Y would be associated. They are unconditionally independent, though, so there cannot be such a connection. *Mutatis mutandis* for Y and a definite cause of X. In addition, we can generalize this rule to situations in which the absence of a causal connection between X and Y is due to a *conditional* independence. This generalization requires more technical machinery, but the central intuition remains the same. Moreover, we can again precisely state the conditions under which the generalized rule can and cannot be applied. The application of the generalized inference rule is complex, but feasible.

[7] For technically minded readers, one must assume that the Markov and Faithfulness conditions hold of the data/generating structure pair, and that there is some single generating structure underlying the two local experiments.

These two inference rules both focus on the absence of a causal connection between two variables in the integrated model. Are we able to say anything at all about when a causal connection *must* appear in the integrated model? There are some situations in which a direct causal connection discovered locally must continue to exist in the integrated model. For example, if $X \rightarrow Y$ is learned locally, and other local models tell us that there are no other causal connections between X (or Y) and any other variable in the integrated model, then $X \rightarrow Y$ must appear in the integrated model. More specifically, if $X \rightarrow Y$ appears in some local model and there is no other *possible* causal connection between X and Y through other variables in the integrated model, then $X \rightarrow Y$ will be part of the integrated model. But in the absence of substantial background knowledge, there is little reason to think that these preconditions will hold with any regularity in practice. We might learn that X causes Y in some local model (whether from experimental or observational data), but that information will typically not enable us to determine whether X is a direct or indirect cause of Y in the integrated model.

Throughout this section, I have avoided talking about functional forms or model parameterizations. For example, I have nowhere assumed that continuous variables are connected (causally) by linear equations. If we can make those types of assumptions, then we can typically learn more about the integrated structure. For example, suppose three different experiments yield models with the following 'causal' connections: *ExposureToInfluenza* → *InfectionWithInfluenza*, and *InfectionWithInfluenza* → *BreathingProblems*, and *ExposureToInfluenza* → *BreathingProblems*. If we know the (unconditional) correlations among these three variables and assume linearity, then although we never directly measure all three variables, we can determine whether *ExposureToInfluenza* and *BreathingProblems* are independent conditional on *InfectionWithInfluenza*. If they are conditionally independent, then the apparent causal connection should be eliminated from the integrated model, even though neither of the above inference rules would license such a removal. I will not further explore inference rules for special functional forms in this paper.

Throughout the above discussion of inference rules, I have mentioned repeatedly that these results all hold only given certain testable assumptions. Interestingly, these inference rules themselves provide a novel means of testing whether these assumptions hold for a particular problem or domain. In theory, the application of these inference rules could yield constraints or predictions that conflict with our background knowledge, or possibly even prior experiments. Since the inference rules (or rather, precise formal statements of them) hold whenever the assumptions do, failure of an inference rule (i.e. production of a false constraint) implies failure of one of the assumptions. I will briefly return to this feature of these results during the application to a case study in Section 4.

I conclude this section by returning to the second and third questions posed in Section 1: how can background knowledge be incorporated, and how can we choose the best sequence of experiments (or heuristically, just the best next experiment) to perform? I have focused throughout this paper on learning causal connections and mechanisms, and so prior knowledge or beliefs could take a variety of forms, including knowledge about the temporal relations among variables, about the definite presence or absence of a particular causal connection, and about the functional form that some connection must have (if it exists). All of these different types of background (i.e. context-dependent) knowledge can be incorporated at two different points in the integration process. First, we can use that background knowledge in whatever method we use for learning local causal models, thereby potentially altering the various models to be integrated. Second, we could use the background knowledge after the application of the inference rules, as a 'post-processor' that excludes some hypotheses from consideration. In this latter mode, the background knowledge could clearly be applied interactively with the inference rules, as implications of the rules might suggest additional (tentative) 'background beliefs' that might further constrain the application of the inference rules.

For the experiment choice problem, a simple naïve algorithm would first enumerate the possible sequences of experiments as well as the possible integration outcomes for each stage in each sequence. We could then apply the above inference rules to each extended experiment-outcome sequence to determine the stage at which we would settle on a unique integrated structure. If we then had some probability distribution over the experiment-outcome sequences, we could determine which experiment sequence has the earliest expected stage at which it settles on a unique model. Of course, this strategy is hopeless from a computational point of view, because it requires both the enumeration of a highly exponential number of sequences and a specification of the probability distribution over experiment-outcome sequences. We can avoid the computational explosion by using some heuristic strategy, but that strategy will not be guaranteed to find the optimal experiment sequence. Unfortunately, we must—in this domain, as in many others—make a decision between asymptotic correctness and computational tractability, and the balancing point for that trade-off depends on the particular domain and scientists. Thankfully, many of these issues have already been explored from a formal perspective in the machine learning community as so-called 'active learning' (see, e.g. Tong and Koller [2001]). That being said, an exploration of the value of different heuristic strategies would constitute a paper in its own right, and so I do not explore it further here.

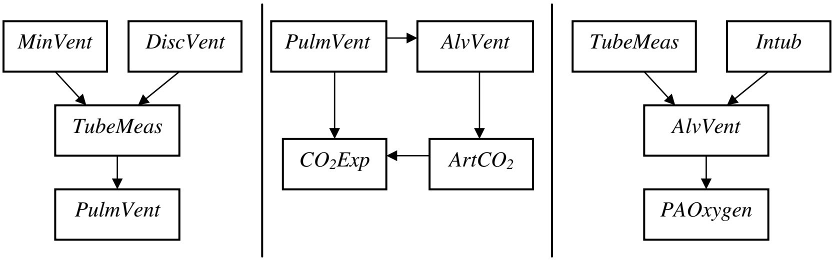

Figure 1. Three local models.

3 A concrete example of the inference rules

To demonstrate in more detail how these inference rules might work, I first present a simplified, but more realistic, example in which a substantial portion of the global structure can be reconstructed from only three instances of local learning. Suppose we are trying to learn a causal model for blood oxygen saturation and ventilators in an intensive care unit. And further suppose that we obtain the three local models given in Figure 1.[8]

We now need to determine what constraints exist on the integrated structure involving the nine variables in these three models. The first inference rule described in the previous section, 'Absence of causal connection in a local model ⇒ absence of causal connection in the integrated model', applies to eight pairs of variables (e.g. between *MinVent* and *DiscVent*), and so we can exclude from the integrated model any causal connections between the variables in those pairs. (A full list of excluded causal connections is given in footnote 9.)

Now consider applying the second inference rule. One consequence of the third local model is that *TubeMeas* and *Intub* must have been unconditionally independent in the data. Therefore, there cannot be a causal connection between *Intub* (or *TubeMeas*) and the definite parents of *TubeMeas* (or *Intub*). Therefore, there cannot be a causal connection in the integrated model between *Intub* and either *MinVent* or *DiscVent*. It is easy to see in this example why there could not be such connections: if there were, for example, an unobserved common cause of *Intub* and *MinVent*, then *Intub* and *TubeMeas* would have been unconditionally correlated (since the first local model tells us that *MinVent* is a cause of *TubeMeas*), and so the third local model would have been different. Also, although the exact details of the

[8] Descriptions of the nine variables: *MinVent*: Minute ventilation measured at the ventilator; *DiscVent*: Disconnected ventilation tube; *TubeMeas*: Ventilation measured at endotracheal tube; *PulmVent*: Pulmonary ventilation; *AlvVent*: Alveolar ventilation; *$ArtCO_2$*: Arterial carbon dioxide content; *CO_2Exp*: Carbon dioxide content of expired gas; *Intub*: Intubation status; *PAOxygen*: Pulmonary artery oxygen saturation.

generalized version of the second inference rule (i.e. the modification to exploit conditional independencies) were left out of the previous section, we can apply the generalized version in this case to exclude more causal connections from the integrated model. For example, *PulmVent* and *ArtCO$_2$* are (by the second model) independent conditional on just *AlvVent*. If there were a causal connection between *ArtCO$_2$* and *TubeMeas* (one of *PulmVent*'s causes), then there would be an unobserved common cause of *PulmVent* and *ArtCO$_2$* relative to the second model, contradicting the observed conditional independence. Therefore, there cannot be any causal connection between *ArtCO$_2$* and *TubeMeas* in the integrated model. Similar applications of the generalized inference rule yield constraints excluding causal connections between four other variable pairs in the integrated model.

In all, there are 15 constraints on the integrated model: that is, there are 15 (unordered) pairs of variables such that there cannot be a causal connection between the two variables in the integrated model.[9] There are 36 unordered pairs of variables that might have some causal connection between them in the integrated model, and so the constraints yielded by these inference rules apply to 42% of the variable pairs in the integrated model. The integrated model is significantly constrained by only three four-variable models. Of course, these results also mean that there are 21 variable pairs about which we must remain agnostic. For example, no constraint is placed on a potential causal connection between *Intub* and *PulmVent*. There is no evidence that suggests such a connection exists, but also no evidence against it. We also have *TubeMeas* → *PulmVent*, *PulmVent* → *AlvVent*, and *TubeMeas* → *AlvVent*, and so might suspect that *TubeMeas*'s influence on *AlvVent* is entirely through *PulmVent*. That is, we might think that *PulmVent* screens off *TubeMeas*'s influence on *AlvVent*. Unfortunately, we cannot determine whether that suspicion is correct based on the local models provided (though we potentially could if we had knowledge of the functional forms).

The example in this section was specifically chosen because we actually have the full integrated model. Beinlich et al. ([1989]) developed a model of the causal interactions among a wide range of variables (37, in all) in an intensive care unit, and the above three local models represent submodels of the full structure. The actual integrated model for these nine variables is shown in Figure 2.

As the full structure shows, all of the constraints are correct; the inference rules worked. Consider also the two variable pairs about which we were

[9] In the interest of completeness, the 15 pairs are: (*CO$_2$Exp*, *AlvVent*); (*MinVent*, *PulmVent*); (*MinVent*, *DiscVent*); (*MinVent*, *ArtCO$_2$*); (*MinVent*, *Intub*); (*MinVent*, *PAOxygen*); (*TubeMeas*, *ArtCO$_2$*); (*TubeMeas*, *Intub*); (*TubeMeas*, *PAOxygen*); (*PulmVent*, *DiscVent*); (*PulmVent*, *ArtCO$_2$*); (*DiscVent*, *ArtCO$_2$*); (*DiscVent*, *Intub*); (*DiscVent*, *PAOxygen*); (*Intub*, *PAOxygen*).

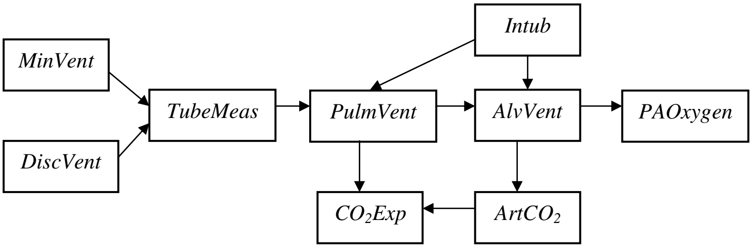

Figure 2. Actual ICU model.

explicitly agnostic: (*Intub*, *PulmVent*) and (*TubeMeas*, *AlvVent*). For the former pair, there turned out to be a (direct) causal connection in the integrated model; for the latter pair, there was not a (direct) causal connection. Our agnosticism turned out to be well-founded.

4 Application to a case study

The strongest test of the applicability of these rules emerges from applying them to a case study in which we do not actually know the full integrated structure. In addition, we want to try to apply the rules to an actual scientific domain and experiments. That being said, bear in mind that we focus throughout this section on what could—normatively—be concluded from a particular study, not on whether scientists actually use these rules in scientific practice. Tuyttens et al. ([1999]) performed an observational study in England to determine which factors causally influence the trappability of badgers (defined as the percentage of individuals trapped during some extended trapping event). Badgers are thought to help spread bovine tuberculosis in England, and so learning the causes of trappability could lead to major shifts in public policy. As a first step in determining the causes of trappability, Tuyttens et al. ([1999]) trapped badgers in three different areas of England, measured potentially causally relevant variables for each trapped badger, and also estimated several population-level variables for each of the three different populations studied.

The principal result of interest here is that badger social group size is (unconditionally) uncorrelated with trappability. Larger groups are no more or less trappable than smaller groups. Thus, if we were to show that the required assumptions hold (a task I do not undertake here), then both inference rules would apply. This independence (i.e. absence of causal connection) fits the precondition of the first rule, and so implies that there cannot be a direct causal connection between trappability and social group size in any integrated model that contains these variables. The second inference rule also applies, in that no integrated model can contain a direct causal connection between trappability (or social group size) and any definite causes of social group size (or trappability) in some other local model. The causes of trappability are unknown (which provided the initial motivation for the Tuyttens et al. [1999] paper), but one potential cause of social group size is the average biomass of earthworms, a major food source for badgers (Johnson et al. [2001]). If earthworm biomass actually is a definite cause of social group size (no definite conclusion is drawn in Johnson et al. [2001]), then the second inference rule implies that there cannot be any direct causal connection between earthworm biomass and trappability in any integrated causal models (given certain testable assumptions).

This implication of the inference rules is perhaps surprising, since a plausible causal story is: 'earthworm biomass is a cause of a badger's hunger, and hunger is a cause of trappability'. Thus, one might have thought a priori that there could be a causal connection between earthworm biomass and trappability. In fact, the causal story just given might be true if one or more of the testable assumptions alluded to earlier is false. Part of the burden of applying these inference rules (and more generally, of doing causal inference) is determining whether the relevant assumptions hold. Alternately, it is possible that the formal framework used to justify the above inference rules is itself not appropriate for the studied variables or domain. And a third possibility is simply that the above a priori causal story is wrong; perhaps the actual model is: 'hunger is a cause of trappability, but reduced earthworm biomass only reduces the group size, rather than increasing an individual badger's hunger'. Or perhaps some other, as yet undetermined, integrated model governs all of these variables.

This pair of articles (Tuyttens et al. [1999]; Johnson et al. [2001]) is somewhat unusual, since the inference rules essentially apply immediately to the results in them. Many (arguably most) individual scientific articles focus either on discovering single causal mechanisms, or on learning the functional form for a particular known causal mechanism. Thus, the value of the inference rules will more frequently emerge when trying to integrate relatively large sets of scientific results (i.e. articles), rather than when focusing on some particular experiment. For example, we might have one scientific paper demonstrating that X is a cause of Y, another showing that Y is a cause of Z, a third showing that W is a cause of Z, and a fourth concluding that there is no causal connection between Y and W. From these four different experiments or papers, we can conclude that there are no causal connections between Y and W or between X and W in any integrated model, but no more. The inference rules provide constraints and guidelines for the development of an integrated model, but there will invariably be questions about the integrated model not answered—because they are not, in fact, answerable—by the information in the local models (e.g. does X directly cause Z?).

I began this paper by arguing that the ultimate goals (and actual end-products) of the various sciences—and particularly applied sciences—include wide-ranging causal structures, but that many different considerations lead us to experiment on only limited sets of variables. Clearly, systematic experimental exploration of all possible subsets of variables is hopelessly impractical. This situation leads to questions about whether local models can be integrated in a normative manner, or whether our global causal models can only be coherent by heuristically piecing them together, perhaps by hand. Perhaps surprisingly, there turn out to be inference rules that enable us to place constraints on the integrated model, and those rules can exploit information that is not contained in any single local model. Novel information can emerge from the consideration of *sets* of local models, as demonstrated by the two different examples. Further applications of these rules to actual scientific practice must await a different paper.

Acknowledgements

Thanks to Maralee Harrell, Clark Glymour, and an anonymous reviewer for helpful comments on an earlier version of this paper.

Department of Philosophy
Carnegie Mellon University
and Institute for Human & Machine Cognition
Pittsburgh, PA 15213
USA
ddanks@cmu.edu

References

Beinlich, I. A., Suermondt, H. J., Chavez, R. M. and Cooper, G. F. [1989]: 'The ALARM Monitoring System: A Case Study with Two Probabilistic Inference Techniques for Belief Networks', in J. R. W. Hunter, J. Cookson and J. Wyatt (*eds.*), 1989, *Proceedings of the Second European Conference on Artificial Intelligence in Medicine*, Berlin: Springer-Verlag, pp. 247–56.

Bessler, D. A. [2003]: 'On World Poverty: Its Causes and Effects', Food and Agricultural Organization (FAO) of the United Nations, Research Bulletin, Rome.

Bovens, L. and Hartmann, S. [2002]: 'Bayesian Networks and the Problem of Unreliable Instruments', *Philosophy of Science*, **69**, pp. 29–72.

Cartwright, N. [1989]: *Nature's Capacities and Their Measurement*, Oxford: Oxford University Press.

Cartwright, N. [1999]: *The Dappled World: A Study of the Boundaries of Science*, Cambridge: Cambridge University Press.

Conati, C., Gertner, A., VanLehn, K. and Druzdzel, M. J. [1997]: 'On-line Student Modeling for Coached Problem Solving using Bayesian Networks', in A. Jameson, C. Paris and C. Tasso (*eds.*), 1997, *Proceedings of the Sixth International Conference on User Modeling* (UM-96), Vienna: Springer-Verlag. pp. 231–42.

Cooper, G. F., Fine, M. J., Gadd, C. S., Obrosky, D. S. and Yealy, D. M. [2000]: 'Analyzing Causal Relationships Between Treating Clinicians and Patient Admission and Mortality in Low-risk Pneumonia Patients', *Academic Emergency Medicine*, **7**, pp. 470–1.

Danks, D. [2002]: 'Learning the Causal Structure of Overlapping Variable Sets', in S. Lange, K. Satoh and C. H. Smith (*eds.*), 2002, *Discovery Science: Proceedings of the 5th International Conference*, Berlin: Springer-Verlag, pp. 178–91.

Danks, D. [2003]: 'Learning Integrated Structure from Distributed Databases with Overlapping Variables', Technical Report CMU-PHIL-149, Department of Philosophy, Carnegie Mellon University.

Danks, D., Griffiths, T. L. and Tenenbaum, J. B. [2003]: 'Dynamical Causal Learning', in S. Becker, S. Thrun and K. Obermayer (*eds.*), 2003, *Advances in Neural Information Processing Systems 15*, Cambridge, MA: The MIT Press, pp. 67–74.

Danks, D., Glymour, C. and Spirtes, P. [2003]: 'The Computational and Experimental Complexity of Gene Perturbations for Regulatory Network Search', in W. H. Hsu, R. Joehanes and C. D. Page (*eds.*), 2003, *Proceedings of IJCAI-2003 Workshop on Learning Graphical Models for Computational Genomics*, pp. 22–31.

Glymour, C. [1999]: 'Rabbit Hunting', *Synthese*, **121**, pp. 55–78.

Glymour, C. [2002]: *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*, Cambridge, MA: The MIT Press.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T. and Danks, D. [2004]: 'A Theory of Causal Learning in Children: Causal Maps and Bayes Nets', *Psychological Review*, **111**, pp. 3–32.

Hesslow, G. [1976]: 'Discussion: Two Notes on the Probabilistic Approach to Causality', *Philosophy of Science*, **43**, pp. 290–2.

Hitchcock, C. [2001]: 'The Intransitivity of Causation Revealed in Equations and Graphs', *The Journal of Philosophy*, **98**, pp. 273–99.

Johnson, D. D., Baker, S., Morecroft, M. D. and Macdonald, D. W. [2001]: 'Long-term Resource Variation and Group Size: A Large-sample Field Test of the Resource Dispersion Hypothesis', *BMC Ecology*, **1**, p. 2.

Lerner, U., Moses, B., Scott, M., McIlraith, S. and Koller, D. [2002]: 'Monitoring a Complex Physical System using a Hybrid Dynamic Bayes Net', in A. Darwiche and N. Friedman (*eds.*), 2002, *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference*, San Francisco, CA: Morgan Kaufmann Publishers, pp. 301–10.

McKim, V. R. and Turner, S. P. [1997]: *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences*, Notre Dame, IN: University of Notre Dame Press.

Nemani, R. R., Keeling, C. D., Hashimoto, H., Jolly, W. M., Piper, S. C., Tucker, C. J., Myneni, R. B. and Running, S. W. [2003]: 'Climate-Driven Increases in Global Terrestrial Net Primary Production from 1982 to 1999', *Science*, **300**, pp. 1560–3.

Pearl, J. [2000]: *Causality: Models, Reasoning, and Inference*, Cambridge: Cambridge University Press.

Ramsey, J., Gazis, P., Roush, T., Spirtes, P. and Glymour, C. [2002]: 'Automated Remote Sensing with Near Infrared Reflectance Spectra: Carbonate Recognition', *Data Mining and Knowledge Discovery*, **6**, pp. 277–93.

Shipley, B. [2000]: *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference*, Cambridge: Cambridge University Press.

Smith, V. A., Jarvis, E. D. and Hartemink, A. J. [2002]: 'Evaluating Functional Network Inference Using Simulations of Complex Biological Systems', *Bioinformatics*, **18**, pp. S216–24.

Spirtes, P., Glymour, C. and Scheines, R. [2000]: *Causation, Prediction, and Search*, 2001, 2nd edn. Cambridge, MA: The MIT Press.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. and Blum, B. [2003]: 'Inferring Causal Networks from Observations and Interventions', *Cognitive Science*, **27**, pp. 453–89.

Tong, S. and Koller, D. [2001]: 'Active Learning for Structure in Bayesian Networks', in B. Nebel (*ed.*), *Proceedings of the International Joint Conference on Artificial Intelligence*, San Francisco, CA: Morgan Kaufmann Publishers, pp. 863–9.

Tuyttens, F. A. M., MacDonald, D. W., Delahay, R., Rogers, L. M., Mallinson, P. J., Donnelly, C. A. and Newman, C. [1999]: 'Differences in Trappability of European Badgers *Meles meles* in Three Populations in England', *Journal of Applied Ecology*, **36**, pp. 1051–62.

Waldemark, J. and Norqvist, P. [1999]: 'In-Flight Calibration of Satellite Ion Composition Data Using Artificial Intelligence Methods', in C. Glymour and G. Cooper (*eds.*), *Computation, Causation, and Discovery*, Cambridge, MA: The MIT Press and AAAI Press, pp. 453–80.

Waldmann, M. R. and Martignon, L. [1998]: 'A Bayesian Network Model of Causal Learning', in M. A. Gernsbacher and S. J. Derry (*eds.*), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, Mahwah, NJ: Erlbaum, pp. 1102–7.

Woodward, J. [2003]: *Making Things Happen: A Theory of Causal Explanation*, Oxford: Oxford University Press.