Comparing Rubin and Pearl's Causal Modeling Frameworks: A Commentary on Markus (2021)

Naftali Weinberger

October 31, 2021

Abstract

Markus (2021) argues that the causal modeling frameworks of Pearl and Rubin are not "strongly equivalent", in the sense of saying "the same thing in different ways". Here I rebut Markus' arguments against strong equivalence. The differences between the frameworks are best illuminated not by appeal to their causal semantics, but rather reflect pragmatic modeling choices.

1 Introduction

Keith Markus' (2021) comparison of the causal frameworks associated with Pearl, Rubin, and Lewis is a gift to scholars of causation. The differences between Pearl and Rubin's frameworks – called structural causal models (SCMs) and Rubin causal models (RCMs), respectively – have been especially obscure to outsiders not already committed to one of them. As each has impacted a wide swath of disciplines (which tend to adopt one or the other) the question of whether they differ in style or substance is significant for causal methodology. Markus' article offers both a guide to those perplexed by the competition between these frameworks and a demonstration that comparing them is philosophically worthwhile. I am hopeful that Markus' article will serve as a starting point for a fruitful literature comparing the approaches, and thus offer this commentary evaluating what it has and has not established.

2 Strong and Weak Equivalence

Although I will focus on SCMs and RCMs, a brief comparison of Pearl and Lewis will help situate Markus' discussion. Lewis (1973) provides counterfactuals with a "possible worlds" semantics. He views possible worlds as conceptually prior to causes, in the sense that he explicates causation using counterfactuals. Galles and Pearl (1998) discuss Lewis' counterfactual semantics in isolation from his philosophical commitments, and prove that were one (contra Lewis) to interpret claims about the closeness of possible worlds in terms of interventions on

variables in causal models, doing so would require no restrictions beyond those already in Lewis' framework. While they do not claim to have shown that Pearl and Lewis' frameworks are equivalent, Pearl does claim this elsewhere (Pearl et al. 2016, p. 126).

Markus argues that Pearl and Lewis' frameworks are not "strongly equivalent" in the sense of saying "the same thing in different ways" (2021, p. 3). At best, Galles and Pearl show that one can take Lewis' notation and assimilate it into Pearl's framework. This might demonstrate "weak equivalence" (3), meaning that one can give the formal expressions in one framework an interpretation within the others. But it does not show that the expressions within each framework express the same things. Of course, Pearl can argue that the ability to do without a possible worlds metaphysics is an advantage of his account (Pearl 2009, § 7.4.1; Woodward 2003, § 3.6), but this advantage does not derive from the frameworks being strongly equivalent. Clearly, they are not.

Markus sees Pearl's comparison of Rubin's framework with his own as flawed in a similar way. Pearl adopts RCM notation to express causal counterfactuals, and interprets these counterfactuals within his own framework. As with Pearl's discussion of Lewis, this strategy can only establish the weak equivalence of the frameworks. This opens the door to asking whether RCMs and SCMs are in fact strongly equivalent, and Markus argues they are not. He also raises concerns about whether they are even weakly equivalent (see Halpern 2000), though here I will focus on his arguments against strong equivalence.

Markus' arguments against strong equivalence highlight ways in which a model within one framework expresses something different than the corresponding model from the other.[1] In my view, this is an unsatisfactory way to evaluate strong equivalence. Since strong equivalence concerns the expressive power of the frameworks, the relevant question is not whether a *particular* model within one framework says the same thing as a model within the other, but rather whether any scenario expressible within one framework can be expressed by *some* (set of) model(s) within the other. Evaluating strong equivalence by pairwise comparison of models amounts to adopting the unreasonably stringent requirement that strongly equivalent theories have a one-to-one correspondence.

More concretely, consider Markus' discussions of correlated disturbances (pp. 7, 11). With SCMs, one either A) assumes the variable set modeled includes all common causes of variables in the set (this is called *causal sufficiency*) or B) uses "bi-directed arcs" to indicate possible unmeasured common causes. Assuming that disturbances (or "error terms") correspond to unmeasured causes of measured variables, this entails that if two variables are not connected by a bi-directed arc, their disturbances have no common cause, and thus will be uncorrelated. Markus emphasizes that within SCMs, accepting a model in which

[1] Markus in fact claims that the *same* model can be interpreted differently within each framework (p. 7), but since models across different frameworks cannot be literally the same, I talk of models as "corresponding to" one another. It will be convenient for the exposition to treat such pairings as unproblematic glosses of what it is for RCM and SCM models to be "the same" – my criticism on Markus depends not on the details of the pairing, but on his assumption that strong equivalence should be tested with pairwise comparisons.

disturbances are uncorrelated amounts to ruling out the possibility that there is a correlation. Such assumptions underwrite results about when a causal effect is identifiable from a probability distribution (Pearl 2009, ch. 3). This contrasts with RCMs, which allow for uncertainty regarding whether disturbances are uncorrelated. Markus presents a scenario in which an SCM model rules out correlated errors, but the corresponding RCM does not, and takes this to show that the frameworks are not strongly equivalent.

Markus sees it as besides the point that SCMs *can* represent correlated disturbances (p. 7), but using a different model than the one he considers. His point is that there are cases where an RCM allows for correlated disturbances, but its SCM counterpart does not. But the fact that some SCM can represent the same scenario as the RCM is what one *should* care about. If it could be shown that any scenario represented in one framework could be represented in the other, this would establish that each framework can say 'the same thing in different ways' and would vindicate Pearl's treatment of the frameworks as inter-translatable. This is not to say that each framework might not be better suited for different aims. Given Pearl's aim of giving a general account of identifiability, it makes sense to design models allowing the user to unambiguously specify that the errors are uncorrelated. To express uncertainty about whether certain error terms are in fact uncorrelated, the SCM modeler could link the relevant variables with a bi-directed arc.[2] But the model does not internally distinguish between the insertion of an arc to indicate belief in the existence of a latent common cause and its insertion to indicate uncertainty, and the ability of RCMs to explicitly represent uncertainty might thus be construed as an advantage. Such pragmatic differences merit philosophical attention, but are not relevant to semantic questions about framework equivalence.

A further argument against strong equivalence relies on the fact that SCMs, but not RCMs, explicitly refer to a causal model in their notation. That is, while Rubin's *potential outcomes* are primitives denoting how individuals would counterfactually respond to experimental treatments, Pearl's counterfactuals are evaluated by reference to a model describing an individual.[3] Markus claims that this rules out strong equivalence. An SCM modeler who adopts a false causal model for an individual will be forced to accept false causal counterfactuals about that individual. In contrast, an RCM modeler can denote counterfactuals about an individual without committing to a particular causal model. Accordingly, SCMs lack the notation to represent a mismatch between one's causal model and the empirical individual one uses it to represent. This, however, does not show that RCMs can represent scenarios that SCMs cannot. In cases where a particular SCM fails to represent an individual, there is an available model that can represent her – namely, the correct model.

Markus' final argument against strong equivalence concerns "non-identical

[2]Note that without assumptions such as minimality, frugality, or faithfulness, variables linked by a bidirected arc are not necessarily correlated.

[3]In table 2, Markus indicates that the counterfactuals in RCMs (but not SCMs) are "model-independent" features of the world. As both frameworks rely on modeling assumption I do not see this distinction as tenable. See Heckman (2005) for relevant discussion.

but necessarily numerically equal" (p. 6) variables. Consider the equation $X = Z$. Interpreted within SCM, this is a "modular" structural equation, meaning that its right-hand side may be replaced while leaving the other equations unchanged. This contrasts with Rubin (1974), who Markus reads as allowing X and Z to necessarily take on the same values. Markus suggests that the SCM with $X = Z$ allows for a wider range of possibilities than the corresponding RCM, as only the former allows the variables to vary independently. Considered, however, in terms of expressive power, this appears to be a further example in which RCMs represent a possibility that SCMs cannot: SCMs cannot represent two variables as both distinct and necessary equivalent.

In what sense are the variables in question "necessarily" equivalent? One possibility is that they are necessarily equal because they denote the same quantity. Since such an equivalence might be non-transparent, one might permit one's framework to represent the variables separately. But this would be a modeler's convenience rather than an extension in the framework's ability to represent states of the world. Another possibility is that the variables refer to distinct quantities that must match due to standing in some non-causal necessitation relationship. Would this prove that RCMs can represent non-causal relationships that SCMs cannot? Not necessarily. A framework can represent causal relationships among a variable set containing non-causally related variables without thereby providing a semantics for the non-causal relationships modelled. Such a framework might allow non-causally related variables into the model, but treat them as a nuisance to be cordoned off to facilitate causal inference. If so, then the framework should not be understood as extending the worldly relationships that can be modeled, but rather as loosening the restrictions on which variables are allowed within causal models.

3 Case Study: Consistency

For the reasons provided, I deny that Markus has shown the frameworks to be not strongly equivalent. Markus would respond that he has, insofar as he has shown that the corresponding models are interpreted differently across the frameworks. The strong/weak distinction is Markus' and he is free to use the terms as he wishes. What matters is whether the distinction supports his critique that when Pearl uses notation from alternative frameworks within his own, it means something different than when interpreted within those frameworks. I have suggested whether formalisms share an interpretation should not be evaluated based on one-to-one correspondence, but rather based on whether the frameworks can express the same causal scenarios. I now motivate this position by appeal to a prior debate between RCM and SCM proponents.

Recall Markus' appeal to the fact that only SCMs explicitly refer to models in their notation. Within SCM, the bridge between models and reality is provided by a theorem called *consistency*. It says that given that a person actually receives a treatment, the observed outcome (i.e. effect) is the one that the model says the individual would have *were* they to receive that treatment (in

SCM notation: $X(u) = x \Rightarrow Y(u) = Y_x(u)$). The status of this principle has been a source of debate between SCM and RCM theorists, and thus serves as a test case for comparisons of the frameworks.[4]

Within Lewis' counterfactual theory (Lewis 1973, § 1.7), consistency follows from the assumption that every world is closest to itself. From the counterfactual "Were I to paint the wall red, my uncle would be happy", it follows that if I actually paint the wall red, my uncle is happy. One might worry that the antecedent could obtain, but with side effects producing an outcome different from that given by the consequent. If the red paint is *toxic*, my uncle's happiness would be a dubious proposition. A consistency defender would reply that if the paint is toxic, one should not accept the stated counterfactual. This back-and-forth regarding consistency is recapitulated among causal modelers (Cole and Frangakis 2009; VanderWeele 2009; Pearl 2010). Consider an SCM licensing the counterfactual that participants in a job program will increase their employability. Yet participants who are *forced* to participate in the program might be resentful and consequently not get its benefits. Pearl's response: if so, then one should *not* accept a model entailing that those individuals would be helped.

Many RCM modelers will not be satisfied with Pearl's response. An experimenter testing the effects of a voluntary job program that does not produce resentment might have no position on whether the program would produce resentment as a side-effect among those who are forced to participate. Yet Pearl's approach requires that if one models the treatment as "job program" and the outcome as "employability", one must take a stance on the general causal relationship between these variables, and thus places a burden on modelers to answer questions they might not want to address. This motivates VanderWeele (2009) to avoid building consistency into the axioms of RCMs, and instead treat it as an empirical assumption requiring case-by-case evaluation.

RCMs interpreted without consistency have fewer implications than the corresponding SCMs. But this is compatible with the frameworks being in an important sense interchangeable. RCM modelers can express the content of SCMs by accepting consistency. And SCM modelers can respond to alleged counterexamples to consistency by providing a model satisfying it. Yet the debate teaches us more than that the frameworks can express the same scenarios. Although it implies the semantic non-equivalence of corresponding SCM and RCM models, at its core it is a pragmatic dispute over modeling methodology. There is a trade-off between requiring more assumption-laden models representing the general relationships among a set of variables and allowing less general models that make fewer commitments, but which are limited to modeling variables within localized experimental contexts. Faced with this trade-off, RCM and SCM modelers make different choices.

[4]Consistency is part of SUTVA (Stable Unit Treatment Value Assumption)(Imbens and Rubin 2015, § 1.6), which is better known for ruling out interaction among units. See Sobel (2006) and Vanderweele et al. (2013) for discussions of how to model such interactions.

4 Manipulability

Although Markus' primary target is the strong equivalence of the frameworks, he briefly considers whether they "assume different forms of causation" (p. 8). His most direct evidence that they do is that Pearl asserts, while some RCM theorists deny,[5] that so-called "non-manipulable" variables can be causes (Pearl 2019; Holland 1986, 2008). Race and gender, which arguably cannot be experimentally manipulated, are key examples of such variables. The disagreement over whether certain variables can be causes suggests that the frameworks make different commitments regarding causation, and is at odds with the more conciliatory treatment of the frameworks I have been defending.

My response is that although advocates of the frameworks adopt conflicting positions regarding certain variables, these positions are not forced upon them by their frameworks. When one moves away from thorny variables such as race and gender and looks at debates regarding slightly less contentious variables such as obesity (Hernán and Taubman 2008; Pearl 2018), the modeling issues in play significantly overlap with those arising in the consistency debate. Whereas RCM modelers link potential outcomes to particular experimental manipulations, SCM modelers represent manipulations by formally applying the do-operator to variables in a graph. Let's reserve the term "interventions" for variables characterized by this operator. Provided that the treatment variable is not "ambiguous" (Spirtes and Scheines 2004), the effects of interventions on an outcome will be invariant across distinct ways of manipulating the treatment. The first-order debate appears to be not over the difference between manipulable and non-manipulable variables, but rather one regarding whether causal claims should be linked to particular manipulations or rather characterized as interventions on variables allowing for distinct manipulations.

Admittedly, Pearl does assert that that one can intervene upon gender without specifying a manipulation. He would, however, require "do(gender)" to be well-defined, which requires there be at least *hypothetical* manipulations on gender (perhaps available only to "Lady Nature herself" (Pearl 2018, p. 4)). Whether such a manipulation is coherent is debatable, and resolving this debate would require careful attention to the purportedly non-manipulable variable. Given SCM modelers' willingness to characterize interventions in a way that abstracts away from concrete manipulations, it is unsurprising that they would have a higher tolerance than RCM modelers for talk of hypothetical manipulations. Yet the frameworks themselves do not settle what one should say about particular "non-manipulable" variables.

5 Individuals and Populations

While I have here focused on Markus' central philosophical thesis, my argument in no way undermines the value of Markus' characterization of the differences

[5]Many methodologists (e.g. VanderWeele 2016) view RCMs as tools for quantitatively defining causal effects, and caution against drawing conclusions for what counts as a cause.

between the approaches, summarized in his table 2 (p. 12). My criticisms only target his explanation of these differences by appeal to the non-equivalence of the frameworks. I further endorse his positions that Pearl has not *established* strong equivalence, and that even if he had, comparing the frameworks would still be worthwhile.

I will now highlight one benefit of considering the potential outcomes framework alongside Pearl's. The former, by including a subscript for the individual in its notation, forces the user to attend to issues of aggregation and abstraction in a way that SCM does not. This is evident from the centrality of "the fundamental problem of causal inference" (Holland 1986) to RCM. The crux of the problem is that although an individual's causal effect is the difference between her outcomes under treatment and control conditions, one only ever observes one of these outcomes. The solution is that, in the limit, randomization ensures that the difference in expected outcomes between the treatment and control groups measures the average effect across the individuals. Note that the average effect is just as much identifiable with the SCM framework, and that the RCM framework never in reality identifies the effect for an individual characterized using a maximally fine-grained description. But the RCM framework makes salient the way that population-level causal relationships aggregate over individual-level effects in a way that may not be transparent when using an SCM to identify the relationships given a joint probability distribution.

One might suppose that RCMs' emphasis on individual-level causes indicates that they interpret causation differently from SCMs. Yet population- and individual-level causes need not be understood as picking out distinct causal concepts. The view that they are is encouraged by the position (Sober 1984) that "type" and "token" causation pick out two metaphysical relations, one between properties and the other between events. Yet careful observers of SCMs have denied that claims about populations and individuals employ distinct metaphysical concepts (Hausman 2005; Woodward 2003, p. 40). Individuals can be considered either as concrete tokens or as types characterized by their properties, and type-level causal relationships generalize over counterfactuals about token individuals.[6] RCM discussions of the "fundamental problem" support this analysis. The view that claims about individuals and populations pick out distinct causal concepts remains prevalent, but in my view should be abandoned.

6 Conclusion

I conclude that Markus' has shown neither that RCM and SCM are strongly non-equivalent nor that they employ distinct notions of causation. Disputes between proponents of the frameworks are better understood as what Weinberger and Bradley (2020) call a "non-factual disagreement". Non-factual disagreements concern not some first-order fact within the domain of dispute, but reflect dif-

[6] Actual causation (Halpern and Hitchcock 2015) is sometimes called "token" causation, but accounts of this notion engage in a project of limited relevance to causal inference and estimation – that of ascribing post-hoc responsibility for the occurrence of an effect.

ferent views of the aims and methods for studying the domain. Regarding SCM and RCM modelers, Markus succinctly captures their distinct aims as follows:

> SCM seeks to encapsulate general scientific knowledge represented in multi-purpose causal models and use them to guide estimation of various causal effects included in the model. In contrast, RCM instead emphasizes the representation of specific events in the context of a specific study. (p. 9)

The dispute arises because proponents of each frameworks see their aims as primary and view the tools of the other as being ill-suited for addressing the questions they view as most important. Should the frameworks turn out to be strongly equivalent, this would not motivate focusing on one framework to the exclusion of another, as there are insights that arise when using each that are less transparent when using the other. But the insights to be gleaned pertain not to metaphysics, but to modeling.

References

Cole, S. R. and C. E. Frangakis (2009). The consistency statement in causal inference: a definition or an assumption? *Epidemiology 20*(1), 3–5.

Galles, D. and J. Pearl (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science 3*(1), 151–182.

Halpern, J. Y. (2000). Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research 12*, 317–337.

Halpern, J. Y. and C. Hitchcock (2015). Graded causation and defaults. *The British Journal for the Philosophy of Science 66*(2), 413–457.

Hausman, D. M. (2005). Causal relata: Tokens, types, or variables? *Erkenntnis 63*(1), 33–54.

Heckman, J. J. (2005). The scientific model of causality. *Sociological methodology 35*(1), 1–97.

Hernán, M. A. and S. L. Taubman (2008). Does obesity shorten life? the importance of well-defined interventions to answer causal questions. *International journal of obesity 32*(3), S8–S14.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association 81*(396), 945–960.

Holland, P. W. (2008). Causation and race. *White logic, white methods: Racism and methodology*, 93–109.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Lewis, D. (1973). *Counterfactuals*. Blackwell Publishers.

Markus, K. A. (2021). Causal effects and counterfactual conditionals: contrasting rubin, lewis and pearl. *Economics & Philosophy*, 1–21.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. (2010). On the consistency rule in causal inference: Axiom, definition, assumption, or theorem? *Epidemiology*, 872–875.

Pearl, J. (2018). Does obesity shorten life? or is it the soda? on non-manipulable causes. *Journal of Causal Inference 6*(2).

Pearl, J. (2019). On the interpretation of do (x). *Journal of Causal Inference 7*(1).

Pearl, J., M. Glymour, and N. P. Jewell (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*(5), 688–701.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association 101*(476), 1398–1407.

Sober, E. (1984). Two concepts of cause. In *PSA: Proceedings of the biennial meeting of the philosophy of science association*, Volume 1984, pp. 405–424. Philosophy of Science Association.

Spirtes, P. and R. Scheines (2004). Causal inference of ambiguous manipulations. *Philosophy of Science 71*(5), 833–845.

VanderWeele, T. J. (2009). Concerning the consistency assumption in causal inference. *Epidemiology 20*(6), 880–883.

VanderWeele, T. J. (2016). Commentary: on causes, causal inference, and potential outcomes. *International journal of epidemiology 45*(6), 1809–1816.

Vanderweele, T. J., G. Hong, S. M. Jones, and J. L. Brown (2013). Mediation and spillover effects in group-randomized trials: a case study of the 4rs educational intervention. *Journal of the American Statistical Association 108*(502), 469–482.

Weinberger, N. and S. Bradley (2020). Making sense of non-factual disagreement in science. *Studies in History and Philosophy of Science Part A 83*, 36–43.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford university press.