NeuroImage xxx (2011) xxx–xxx

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Comments and Controversies

Cloak and DAG: A response to the comments on our comment

Martin A. Lindquist *, Michael E. Sobel

Department of Statistics, Columbia University, USA

ARTICLE INFO

Article history:
Received 19 September 2011
Revised 3 November 2011
Accepted 8 November 2011
Available online xxxx

ABSTRACT

Our original comment (Lindquist and Sobel, 2011) made explicit the types of assumptions neuroimaging researchers are making when directed graphical models (DGMs), which include certain types of structural equation models (SEMs), are used to estimate causal effects. When these assumptions, which many researchers are not aware of, are not met, parameters of these models should not be interpreted as effects. Thus it is imperative that neuroimaging researchers interested in issues involving causation, for example, effective connectivity, consider the plausibility of these assumptions for their particular problem before using SEMs. In cases where these additional assumptions are not met, researchers may be able to use other methods and/or design experimental studies where the use of unrealistic assumptions can be avoided. Pearl does not disagree with anything we stated. However, he takes exception to our use of potential outcomes' notation, which is the standard notation used in the statistical literature on causal inference, and his comment is devoted to promoting his alternative conventions. Glymour's comment is based on three claims that he inappropriately attributes to us. Glymour is also more optimistic than us about the potential of using directed graphical models (DGMs) to discover causal relations in neuroimaging research; we briefly address this issue toward the end of our rejoinder.

Statistical methods are extremely useful for studying relationships among variables. Additional assumptions, of which researchers are often unaware, are needed to justify interpreting these relationships as indicative of causation, and if these assumptions are not met, the resulting causal inferences will generally be invalid. In our comment (Lindquist and Sobel, 2011), we used potential outcomes' notation (Neyman, 1923 [1990]), which is the standard notation used by statisticians working in the area of causal inference, to explicate additional assumptions that suffice for the parameters of a directed graphical model (DGM) or structural equation model (SEM) to be interpreted as causal effects. These assumptions are in addition to those made when an SEM or DGM is used for descriptive or predictive purposes. As evidenced by their comments, neither Pearl (Pearl, 2011) nor Glymour (Glymour, 2011) are enamored of this notation. Unfortunately, neither says anything about the point that we used this notation to address, namely the need to be explicit about the assumptions that are made when SEMs (DGMs) are used to make causal inferences, in order to assess when such assumptions are and are not plausible in neuroimaging research. We return to this, our main point, after addressing, respectively, the specifics raised by Pearl and Glymour.

Specifics

Pearl

Graphical models have proven to be tremendously useful in a number of application areas (Jordan, 2004). Our concern herein is solely with the use of DGM's for making inferences of a special nature, that is, inferences about causation. In our original comment, following Robins (2003), we exposited, using potential outcomes' notation, the assumptions underlying Robins (1986) causal model for directed acyclic graphs (DAGs). Robins (1986) called this model the "finest fully randomized causally interpreted structured tree graph". As Robins (2003) showed (see his lemma 1), the assumptions underlying the causal interpretation of the nonparametric structural equation model Pearl subsequently put forth are even stronger. Therefore, the assumptions that we described for giving a causal interpretation to a DGM apply as well to Pearl's nonparametric structural equation model, as it is a special case of Robins (1986) model.

To keep matters simple, we took up the DGM corresponding to the DAG $Z \rightarrow X \rightarrow Y$, with Z binary, $Z = 0$ indicating assignment to the control group (and no receipt of treatment) , $Z = 1$ indicating assignment to the treatment group (and receipt of treatment). Under the DGM, the distribution of (Y, X, Z) factorizes as $f(y|x)f(x|z)f(z)$, as the absence of a directed arrow from Z to Y means that Y is independent of Z, given X. We stressed that additional assumptions were needed before estimates computed from DGMs should be given a causal interpretation. We made the following assumptions (see our original comment or

* Corresponding author at: 1255 Amsterdam Ave, Columbia University, New York, NY 10027, USA.
E-mail address: martin@stat.columbia.edu (M.A. Lindquist).

1053-8119/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2011.11.027

Hernán and Robins (forthcoming, 2012) for an introduction to potential outcomes' notation):

(1) the existence of the potential outcomes $X(z)$ and $Y(z, x)$ for all z and x,
(2) $Y(0, x) = Y(1, x)$, expressing the idea that Z does not directly cause Y,
(3) $X = X(Z)$, $Y(z) = Y(z, X(z))$, $Y = Y(Z, X(Z))$,
(4a) $Y(z,x), X(z) \underline{\perp\!\!\!\perp} Z$ for all z, x,
(4b) $Y(z,x) \underline{\perp\!\!\!\perp} X|Z$ for all z, x.

Our point was that by examining these assumptions, researchers can decide whether or not they are substantively reasonable in the particular applications they have in mind. Under assumption 1, each observation has two potential intermediate outcomes, $X(0)$ in the absence of treatment, and $X(1)$ under treatment. It is important to note that this does not necessarily mean we can force cases to receive treatment or not; it is enough to imagine that each case could have received treatment or not. For example, we can speak of the effect of a volcano erupting even if we cannot intervene to make the volcano erupt. Similar remarks apply to the potential outcomes $Y(z, x)$. Assumption 2 states that Z does not directly cause Y. Assumption 3 links the observed values and the potential outcomes. First, the observed value $X = (1 - Z)X(0) + ZX(1)$, reflecting the fact that we observe $X(0)$ when $Z = 0$ and $X(1)$ when $Z = 1$. Second, there are two potential outcomes: $Y(0, X(0))$ in the absence of treatment, and $Y(1, X(1))$ under treatment. Third, the observed value $Y = (1 - Z)Y(0, X(0)) + ZY(1, X(1))$. Finally, assumptions 4a and 4b are critical. 4a states essentially that assignment to the treatment or control group behaves as if randomized, and in many fMRI studies subjects are randomized. 4b states that within the treatment (control) group, assignment to the mediating variable X behaves as if randomly assigned. But here, as assignment to X is not actually randomized, one has to ask whether or not it is reasonable to assume that within the treatment (control) group, one can treat X as if it were. In our original comment, we gave an example of a neuroimaging study where such an assumption would not be reasonable because there was a variable confounding the X–Y relationship. In this case, causal inferences about direct and indirect effects' of Z on Y and effects of X on Y using the SEM (DGM) corresponding to the $Z \rightarrow X \rightarrow Y$ DAG will be incorrect. Thus, we emphasized that researchers need to understand the assumptions they make when a DGM is viewed as a causal model, and that they assess the plausibility of these assumptions in their particular setting before proceeding (or not) to use a DGM to make causal inferences.

Pearl states that assumptions 1–4b can be "derived" from a "counterfactual reading" of the "causal chain" $Z \rightarrow X \rightarrow Y$, and he claims that we are trying to "replace or discredit" this chain. Leaving aside the fact that a DAG cannot be discredited, assumptions 1–4b constitute a specification of the causal model that the DAG is intended to represent. From our point of view, whether or not a researcher wants to draw a picture to represent 1–4b is not even an issue. What is important is that a researcher who does so, thereby using the DAG as a shorthand, understand that assumptions 1–4b constitute the model and that when these assumptions are not met, but the model is nonetheless used, erroneous causal inferences will result. For a complementary perspective, see Dawid's (2009) paper "Beware of the DAG".

Pearl then proceeds toward criticism of potential outcomes' notation, giving his own "reading" (specification of the mathematical model the DAG represents) which does not use this notation. According to Pearl the full graph would specify:

(P1) Y is determined by X only,
(P2) X is determined by Z only,
(P3) All functional relationships are further modified by omitted factors (not shown explicitly in the graph) that are assumed to be mutually independent yet arbitrarily distributed.

Pearl claims that no additional assumptions are required to derive all the conclusions obtained using our assumptions 1–4b, and proceeds to claim that his rendition of the chain is superior to ours ("more transparent, rigorous, explicit and conducive to meaningful scientific discourse"), that potential outcomes' notation is "opaque", and even that more valid results are obtained when counterfactual language is avoided.

At first glance Pearl's formulation appears simple and clear. However, this is illusory, as statements P1–P2 lack meaning, due to the vagueness of the word "determined". Usually when we talk about determination using random variables, we are referring to prediction (stochastic dependence and independence). Under this interpretation, P1 says that $f(y|x, z) = f(y|x)$. Further, that is the mathematical constraint imposed on the joint distribution of (Z, Y, X) when the DGM corresponding to the DAG above is used, and it is the interpretation that most knowledgeable users of SEM's would give. However, it turns out that this is not the intended interpretation. Second, P3 says that the DAG $Z \rightarrow X \rightarrow Y$ is actually an incomplete representation of the actual DAG needed, requiring inclusion of error terms and the specification of probabilistic dependencies among them. In short, P1–P3 are not so clear and simple after all.

Next, Pearl states that assumptions 1–4b can be derived from the DAG, citing one of his papers (Pearl, 2010, pp. 126–127) for "explicit derivation". There, instead of a derivation, we simply found a translation into potential outcomes' notation:

(1) Exclusion restrictions: For every variable Y having parents PA_Y and for every set of endogenous variables S disjoint of PA_Y, we have $Y_{pa_Y} = Y_{pa_Y, s}$.
(2) Independence restrictions: If $Z_1, \ldots, Z_k$ is any set of nodes not connected to Y via dashed arcs, and $PA_1, \ldots PA_k$ their respective set of parents, we have $Y_{pa_Y} \underline{\perp\!\!\!\perp} \{Z_{1pa_1}, \ldots, Z_{kpa_k}\}$.

For the DAG under consideration here, the independence restrictions are essentially our 4a and 4b. The exclusion restrictions, which translate P1 and P2 into potential outcomes' notation, are just our assumption 2. Here, placing things into potential outcomes' notation has a) clarified the use of the word "determined", e.g., Y is determined by X actually means $Y_{zx} = Y_x$ (or $Y(0, x) = Y(1, x)$ in our case), and b) revealed another source of confusion in P1 (and P2), as Y in P1 does not even refer to the observed variable Y (even though this is the natural interpretation), but rather to the potential variable Y_{zx} ($Y(z, x)$ in our notation); in contrast, the relationship between observed and potential outcomes is spelled out in our assumption 3. With the above in mind, we invite the reader to compare assumptions 1–4b with P1–P3 and reassess "which notational system is more transparent…".

More on potential outcomes

Potential outcomes' notation dates back to the work of Neyman (1923, [1990]), and was rediscovered by Rubin (1974). The notation is used to clearly define causal parameters of interest, and to do so independently of the methods used to estimate these, enabling assessment of the assumptions that need to be satisfied in order that an estimation method yields consistent (or unbiased) estimates of the causal parameters of interest. Statisticians have successfully used this notation to clarify the assumptions upon which various long-standing statistical procedures rest; for a few examples, see the work of Holland (1988) on direct and indirect effects (and more generally, mediation) in path analysis, a special kind of SEM, as well as the subsequent work by Robins and Greenland (1992) on direct and indirect effects, the work of Angrist et al. (1996) on instrumental variables, and the work of Robins (1986, 2003) on DGMs. Alternative approaches to mediation (see Frangakis and Rubin (2002) on principal stratification), and a literature on longitudinal causal inference (see Robins and Hernán (2009) for a nice overview) which use potential outcomes' notation have also emerged in recent years, and we

Please cite this article as: Lindquist, M.A., Sobel, M.E., Cloak and DAG: A response to the comments on our comment, NeuroImage (2011), doi:10.1016/j.neuroimage.2011.11.027

suspect that some of these new approaches would not have been developed in the absence of this notation.

Personally, we find that using this notation helps us to formulate problems clearly and avoid making mistakes, to understand and develop identification conditions for estimating causal effects, and, very importantly, to discuss whether or not such conditions are plausible or implausible in practice (as above). Though quite intuitive, the notation requires a little getting used to, primarily because it is not typically included in early statistical training, but once that is accomplished, the notation is powerful and simple to use. Finally, as a strictly pragmatic matter, the important papers in the literature on causal inference (see especially papers by the 3R's (Robins, Rosenbaum, Rubin, and selected collaborators)) use this notation, making an understanding of it a prerequisite for any neuroimaging researcher who wants to learn more about this subject.

Glymour

Glymour makes a number of incorrect statements about potential outcomes and the approach to causal inference that stems from the statistical literature. As a full discussion would require a lengthy response, we address a) only the three claims that Glymour incorrectly attributes to us and b) his assertions that suggest the "potential outcomes framework" is ill-suited for fMRI research.

We begin with the second of Glymour's attributions, namely that we claim "the theory of graphical causal models developed by Spirtes et al. (1993) makes no counterfactual claims". What we actually said (p. 335) was that the model of Spirtes et al. (2000) "does not refer to counterfactuals". Our statement is correct and in no way implies that the model cannot be used to discuss counterfactual claims. Glymour's attribution stems from a failure to distinguish between the definition of a model and the uses to which it may be put. In a more familiar guise, a regression model can be used to make predictions, but the model itself consists of a set of mathematical statements that make no reference to prediction.

Glymour's third attribution appears to suggest we are claiming that observational studies cannot be used to make valid causal inferences. Not at all. The conditions needed to interpret associations as effects are the same in experimental and observational studies. The hypothetical example of ours to which Glymour refers is a randomized experiment in which subjects are assigned to a stress task or a control task. There is an intermediate outcome consisting of the brain response in a stress related area and a final outcome (task performance). Contrary to Glymour, the estimand of interest in our example is the population effect of the brain response on task performance $E(Y(1,1)-Y(1,0))$ under the stress condition, and we simply constructed a substantively plausible example in which one of the sufficient conditions for the SEM (DGM) parameters to be used to make valid causal inferences was violated.

The first (and most general) claim that Glymour misattributes to us is that "'causal effects' cannot be found by methods associated with a variety of directed graph representations of causal relations". The point of our example, and, more generally, our comment, was that when SEMs and related methods are used to make causal inferences, additional assumptions above and beyond those required for using these methods for other purposes (for example, description and prediction) are needed. When these additional assumptions hold, SEMs and related methods provide valid causal inferences. In general, when these assumptions do not hold, causal inferences obtained using SEMs and related methods will be invalid.

Connectivity and search

In the neuroimaging literature, interest in connectivity studies, which describe how various brain regions interact and how these interactions depend on experimental conditions, has increased in recent years. If these studies are properly conducted and the resulting data properly analyzed, they can substantially increase our understanding of how the brain functions and processes information. In the example we presented in our original comment, there is an experimental variable, a mediator and a final outcome, and the goal was to estimate how the effect of a treatment on a behavioral response is mediated by activation in a particular brain region. This is similar to the setup in effective connectivity studies, where one aim is to estimate how the effect of a treatment on activity in one brain region is mediated by activity in a different region. For expository purposes we tried to keep things as simple as possible, using a binary treatment variable and intermediate and final responses that are measured without error. Glymour suggests that what he calls the "PO framework" is, among other things, inherently limited to similar types of situations. That is not true. It is trivial to define potential outcomes for continuous and multivariate treatments, and there is nothing to preclude the response from being defined as a latent variable. The case of resting state studies is also handled readily, and it is no problem to work with more than 3 variables, for example, multiple mediators and/or multiple outcomes. That said, additional identifying assumptions would be required in the case of multiple mediators. As an example, Lindquist (2011) considers the case of a binary treatment, a functional mediator and a subsequent outcome.

At the beginning of our original comment we said (p. 335) that we were "unconvinced that directed graphical models (DGMs) are generally useful for finding 'causal relations' or estimating causal effects". We focused on the latter issue, in part because in studies of effective connectivity, estimation of causal effects holds center stage. However, Glymour seems to be more concerned with discovering causal relations. Our skepticism concerning the use of DGMs for this purpose stems primarily from two considerations. First, like many others, we do not believe that causal relations in fMRI or other areas can be reliably discovered without making substantive assumptions about possible causal relationships (including the case where a variable does not affect another) among variables; see for example Robins and Wasserman (1999) for a similar point of view as well as an explicit critique of search methods. Second, our original comment establishes that conditional associations in DGMs do not evidence causation unless additional assumptions are met. This applies equally well in the context of search using DGMs and we find it difficult to believe that these additional assumptions will be met in all circumstances of interest to neuroscientists.

Where's the science?

Our original note had two aims. First, we wanted neuroimaging researchers to recognize that when they use SEMs to make causal inferences, the validity of the conclusions rest on assumptions above and beyond those required to use an SEM for descriptive or predictive purposes. Unfortunately, these assumptions are rarely made explicit, and in many instances, researchers are not even aware that they are needed. Since these assumptions can have a major impact on the "findings", it is critical that researchers be aware of them, and even though they may not be testable, that they think carefully about the science behind their problem and utilize their substantive knowledge to carefully consider, before using an SEM, whether or not these assumptions are plausible in the particular problem under consideration. For this reason, we were disappointed that Pearl, who in his comment does not disagree with anything we stated about the importance of assumptions, nevertheless ends his comment by urging fMRI researchers "to continue using their familiar SEM language" (and presumably SEMs). This recommendation, which completely ignores the importance of assumptions for causal inference and which is put forth with absolutely no consideration for fMRI subject matter knowledge and considerations that might inform when SEMs are reasonable or not to use, will prove harmful to the field if

Please cite this article as: Lindquist, M.A., Sobel, M.E., Cloak and DAG: A response to the comments on our comment, NeuroImage (2011), doi:10.1016/j.neuroimage.2011.11.027

taken seriously. For if fMRI researchers continue to use their "familiar" approach, drawing diagrams and fitting SEMs without realizing the assumptions they are making, many of the causal inferences thereby generated will be incorrect, and the development and use of alternative ways of studying effective connectivity will be stifled. That was our second aim: to encourage researchers to develop and use such alternative procedures (for example, instrumental variables and principal stratification) when SEMs and related methods are inappropriate. Additionally, we wanted to encourage researchers to develop alternative experimental designs that might be used to identify both direct effects of mediating variables and indirect effects of treatments on outcomes. As an example, we pointed to the use of combining fMRI and transcranial magnetic stimulation, using the latter technique to randomize subjects on the mediator, thereby enabling inferences using standard procedures, including SEM's.

As an example illustrating several of the points in the preceding paragraph, consider the following hypothetical randomized experiment using a social evaluative threat task (Wager et al., 2009). A researcher is interested in performing an effective connectivity analysis where Z is a binary treatment assignment variable (task/no task), and X and Y are, respectively, the observed BOLD responses in ventromedial prefrontal cortex (VMPFC) and brainstem periaqueductal gray (PAG) regions. The researcher hypothesizes that treatment causes activity in VMPFC, that activity in VMPFC causes activity in PAG, and that there is no "direct effect" of the treatment on activity in PAG (our assumptions 1 and 2). Assuming this is correct, population level causal effects can be estimated using the SEM we have been considering throughout if assumptions 4a and 4b are met. Since subjects are randomized to the treatment or control group, 4a holds. However, suppose there is some latent behavioral or genetic trait, for example anxiety, such that subjects with the trait tend to have a higher response in both VMPFC and PAG compared to those without the trait. Assumption 4b is then violated and the SEM we have been considering would not provide a valid estimate of the effect of activity in VMPFC on activity in PAG. Sobel (2008) shows that the instrumental variable estimand can be used to correctly estimate this effect in a two equation linear system under assumptions less restrictive than 4b (that is, if 4b holds, these alternative assumptions hold). A researcher can then examine these alternative and weaker assumptions and ask whether these might plausibly hold in the application above.

It should be noted that this example is slightly different than most standard effective connectivity examples where a single group of subjects is observed under different experimental conditions. In addition, we do not directly address the temporal aspect of the connectivity problem and issues related to the variation in hemodynamic response delays across regions of the brain (e.g., Lindquist and Wager, 2007; Lindquist et al. 2009). However, the ideas presented herein can be extended to these settings as well, something we shall address in future work. Further, although this comment focuses on SEMs, the issues discussed in the example and more generally in our commentary also apply to other commonly used techniques for assessing effective connectivity (e.g., Granger causality, DCM, Bayes Net).

In short, causal inferences are only as valid as the assumptions upon which they rest. Paying careful attention to assumptions and using appropriate methods to make inferences about causation moves science forward. Dogmatic adherence to a particular methodology moves it rearward.

Acknowledgments

We would like to thank Bengt Muthen, Tor Wager and several anonymous reviewers for their helpful comments on the manuscript.

References

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. J. Am. Stat. Assoc. 91, 444–455.
Dawid, A.P., 2009. Beware of the DAG! J. Mach. Learn. Res. Workshop Conf. Proc. 6, 59–86.
Frangakis, C.E., Rubin, D.B., 2002. Principal stratification in causal inference. Biometrics 58, 21–29.
Glymour, C., 2011. Counterfactuals, graphical causal models and potential outcomes: Response to Lindquist and Sobel. NeuroImage doi:10.1016/j.neuroimage.2011.07.071.
Hernán, M.A., Robins, J.M., 2012. Causal Inference. Chapman Hall/CRC.
Holland, P.W., 1988. Causal inference, path analysis and recursive structural equation models (with discussion). In: Clogg, C.C. (Ed.), Sociological Methodology 1988. American Sociological Association.
Jordan, M.I., 2004. Graphical models. Stat. Sci. 19, 140–155.
Lindquist, M.A., 2011. Functional Causal Mediation Analysis with an application to Brain Connectivity. In submission.
Lindquist, M.A., Sobel, M.E., 2011. Graphical models, potential outcomes and causal inference: comment on Ramsey, Spirtes and Glymour. NeuroImage 57, 334–336.
Lindquist, M.A., Wager, T.D., 2007. Validity and power in hemodynamic response modeling: a comparison study and a new approach. Hum. Brain Mapping 28, 764–784.
Lindquist, M.A., Loh, J.M., Atlas, L., Wager, T.D., 2009. Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. NeuroImage 45, S187–S198.
Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9, translated in *Statistical Science*, (with discussion) 5, 465–480 1990.
Pearl, J., 2010. The foundations of causal inference. In: Liao, T. (Ed.), Sociological Methodology 2010. Wiley.
Pearl, J., 2011. Graphical models, potential outcomes and causal inference: Comment on Lindquist and Sobel. NeuroImage 58, 770–771.
Robins, J.M., 1986. A new approach to causal inference in mortality studies with sustained exposure periods — application to control of the healthy worker survivor effect. Math. Model. 7, 1393–1512.
Robins, J.M., 2003. Semantics of causal DAG models and the identification of direct and indirect effects. In: Green, P., Hjort, N., Richardson, S. (Eds.), Highly Structured Stochastic Systems. Oxford University Press.
Robins, J.M., Greenland, S., 1992. Identifiability and exchangeability of direct and indirect effects. Epidemiology 3, 143–155.
Robins, J.M., Hernán, M.A., 2009. Estimation of the effects of time-varying exposures. In: Fitzmaurice, G., Davidian, M., Verbeke, G., Molenberghs, G. (Eds.), Longitudinal Data Analysis. Chapman and Hall/CRC.
Robins, J.M., Wasserman, L., 1999. On the impossibility of inferring causation from association without background knowledge. In: Glymour, C., Cooper, G. (Eds.), Computation, Causation and Discover. AAAI Press/MIT Press.
Rubin, D.B., 1974. Estimating causal effects of treatment in randomized and nonrandomized studies. J. Educ. Psychol. 66, 688–701.
Sobel, M.E., 2008. Identification of causal parameters in randomized studies with mediating variables. J. Educ. Behav. Stat. 33, 230–251.
Spirtes, P., Glymour, C., Scheines, R., 1993. Causation, Prediction and Search, 1st Edition. MIT Press.
Spirtes, P., Glymour, C., Scheines, R., 2000. Causation, Prediction and Search, 2nd Edition. MIT Press.
Wager, T.D., Waugh, C.E., Lindquist, M.A., Noll, D.C., Fredrickson, B.L., Taylor, S.F., 2009. Brain mediators of cardiovascular responses to social threat, Part I: reciprocal dorsal and ventral sub-regions of the medial prefrontal cortex and heart-rate reactivity. NeuroImage 47, 821–835.

Please cite this article as: Lindquist, M.A., Sobel, M.E., Cloak and DAG: A response to the comments on our comment, NeuroImage (2011), doi:10.1016/j.neuroimage.2011.11.027