Published in final edited form as:

Neuroimage. 2011 July 15; 57(2): 334–336. doi:10.1016/j.neuroimage.2010.10.020.

Graphical Models, Potential Outcomes and Causal Inference: Comment on Ramsey, Spirtes and Glymour

Martin A. Lindquist* and Michael E. Sobel

Department of Statistics, Columbia University

Abstract

Ramsey, Spirtes and Glymour (RSG) critique a method proposed by Neumann et al (2010) for the discovery of functional networks from fMRI meta-analysis data. We concur with this critique, but are unconvinced that directed graphical models (DGMs) are generally useful for estimating causal effects. We express our reservations using the “potential outcomes” framework for causal inference widely used in statistics.

We concur with Ramsey, Spirtes and Glymour’s (RSG) critique of Neumann et al (2010), but are unconvinced that directed graphical models (DGMs) are generally useful for “finding causal relations” or estimating causal effects. As fMRI researchers are more familiar with recursive linear structural equation models (RLSEMs), which are essentially DGM’s with additional structure, we also refer to these throughout. We express our reservations using the “potential outcomes” framework for causal inference widely used in statistics. We briefly introduce this framework here. See Sobel (2009) for a more extended introduction.

Causal relationships sustain counterfactual conditional statements. For example, the statement “John lived because he received treatment” means John took treatment and lived, and would have died otherwise. To formally represent such statements, statisticians developed “potential outcomes” notation. Let $Y_i(0) = 1$ if subject i lives without taking treatment, 0 otherwise; let $Y_i(1) = 1$ or 0 denote these outcomes when treatment is taken. For each subject, the unit causal effect $Y_i(1) - Y_i(0)$ compares i ’s potential outcomes under the two conditions. This effect cannot be observed as we can only see at most one outcome per subject. Nevertheless, under additional conditions, the average treatment effect (ATE) $E(Y(1) - Y(0))$ can be consistently estimated from sample data. The observed data are (Z_i, Y_i) , $i = 1, \dots, n$, where $Z_i = 1$ if unit i receives treatment, 0 otherwise, and $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$ is i ’s observed outcome. To estimate the ATE, the difference between sample means, $\bar{Y}_{\{Z=1\}} - \bar{Y}_{\{Z=0\}}$ is often used. But this estimates

© 2010 Elsevier Inc. All rights reserved.

*Corresponding author: Columbia University, 1255 Amsterdam Ave, New York, NY 10027, martin@stat.columbia.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

$$E(Y|Z=1) - E(Y|Z=0) = E(Y(1)|Z=1) - E(Y(0)|Z=0), \quad (1)$$

and in general (1) = ATE. Thus, a non-zero value of (1) implies Z and Y are associated, but not that Z causes Y . However, if treatment is marginally “ignorable”, i.e., $Y(z) \perp\!\!\!\perp Z$ for $z = 0, 1$ (where $\perp\!\!\!\perp$ denotes independence), as in a randomized experiment, (1) = ATE and association is causation.

To illustrate, consider the hypothetical experiment in Table 1. Because the unit effects cannot be observed, the ATE must be estimated by comparing subjects observed under different conditions. Here $\alpha = E(Y|Z=1) - E(Y|Z=0) = 1.9 - 0.5 = 1.4$. Whether or not $\alpha = \text{ATE}$ depends on the missing potential outcomes in Table 1. If these are missing completely at random (equivalently, treatment is marginally ignorable), then $\alpha = \text{ATE}$. Thus, in a randomized experiment $\bar{Y}_{\{Z=1\}} - \bar{Y}_{\{Z=0\}}$ is an unbiased (and consistent) estimator of the ATE.

Turning attention to the methods described in RSG, a DGM represents the conditional independence/dependence structure among a set of variables. This structure can be visually represented using directed acyclic graphs (DAGs), where the absence/presence of a directed edge has a precise probabilistic meaning. Similarly, path diagrams are used to represent RLSEMs; here the absence/presence of a path represents a 0 coefficient.

Empirical researchers do not use path diagrams or DAGs to represent the structure of association among variables per se, but to represent substantive causal hypotheses. An RLSEM (or DGM) is then used to make causal inferences. As above, the validity of such inferences, which equate association and causation, rest upon additional assumptions that are not part of the RLSEM or DGM. Often these assumptions are neither explicitly specified nor substantively reasonable.

To illustrate the kind of additional assumptions needed for causal inferences of the type above to be sustained using DGMs we consider the “finest fully randomized causally interpreted structured tree graph” due to Robins (1986, 2003). In addition to linking DGM’s with potential outcomes notation, Robins (2003) also discusses other “causal DGM’s”, most notably the so called agnostic causal model of Spirtes, Glymour and Scheines (2000), which does not refer to counterfactuals, and Pearl’s (1995) non-parametric structural equation model, a minor variation on Robins (1986).

To keep matters at their simplest, Figure 1 depicts a DAG and corresponding DGM for three variables Z, X, Y , where Z is binary, taking values 0 and 1. To construct a causal model equivalent to the DGM, in which Z “directly causes” X , X “directly causes” Y and Z does not “directly cause” Y , Robins assumes

- (1) the existence of the potential outcomes $X(z)$ and $Y(z, x)$ for all z and x ,
- (2) $Y(0, x) = Y(1, x)$ for all 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) $Y(z, x), X(z) \perp\!\!\!\perp Z$ for all z, x ,

$$(4b) Y(z, x) \perp\!\!\!\perp X | Z \text{ for all } z, x.$$

Under these assumptions, the distribution $f(x, z, y)$ factors as $f(y | x)f(x | z)f(z)$ as in the DGM corresponding to the DAG. Inferences about the effect of Z on X involve comparison of $f(x | Z = 1)$ with $f(x | Z = 0)$. By (1) (3) and (4a), $f(x | Z = z) = f(x(z))$. Similarly, inferences about the effect of X on Y at x versus x^* involve comparison of $f(y | x) = f(y(0, x)) = f(y(1, x))$ with $f(y | x^*) = f(y(0, x^*)) = f(y(1, x^*))$.

In a randomized fMRI experiment with a treatment and a control group, the potential outcomes $Z(0)$ and $Z(1)$ are well defined, but it is unclear how values of X in $Y(z, x)$ are set. Some authors have even argued that as X is not manipulated in experiments where Z is randomly assigned, potential outcomes $Y(z, x)$ should not be considered. Assumption (4a) states that treatment is ignorable, as would be the case in a randomized fMRI experiment. Assumption (4b) states that potential outcomes $Y(z, x)$ are ignorable with respect to the intermediate outcomes $X(z)$ given Z , as would be the case if subjects were randomly assigned to X at both levels of Z .

Even when assumption (4a) holds, assumption (4b) is very strong, and it is easy to construct substantively plausible examples where this assumption is violated. Suppose subjects are randomized to perform either a stress ($Z = 1$) or control ($Z = 0$) task. There are two types of subjects, resilient ($T = a$) and non-resilient ($T = b$), with probabilities π_a and π_b , respectively. Subject type is an attribute determined prior to randomization; thus, the following stronger version of (4a) holds:

$$(4a') Y(z, x), X(z), T \perp\!\!\!\perp Z \text{ for all } z, x.$$

The intermediate outcome X is the brain response in a key stress-related area of the brain, and Y is task performance. Suppose for simplicity that X is binary, though this assumption can be relaxed, with $\Pr(X = 1 | Z = 1, T = a) = .75$, $\Pr(X = 1 | Z = 1, T = b) = .25$. Suppose also that resilient subjects perform better on average than non-resilient subjects at every level of X under either treatment, implying $\mu_a(1, x) = E(Y(1, x) | T = a) > \mu_b(1, x) = E(Y(1, x) | T = b)$. Finally, suppose

$$E(Y(1, x) | X = 1, T = a) = E(Y(1, x) | T = a), \quad E(Y(1, x) | X = 1, T = b) = E(Y(1, x) | T = b).$$

Then

$$E(Y(1, x) | X = 1, Z = 1) = E(Y(1, x) | X(1) = 1) = \frac{.75\pi_a\mu_a(1, x) + .25\pi_b\mu_b(1, x)}{.75\pi_a + .25\pi_b}.$$

In contrast, $E(Y(1, x)) = \pi_a\mu_a(1, x) + \pi_b\mu_b(1, x)$. In general, $E(Y(1, x)) \neq E(Y(1, x) | X(1) = 1)$. Thus assumption (4b) is violated. In other words, subject type is not a confounder for the relationship between treatment and the outcomes X and Y because the experiment is randomized, but subject type is a confounder for the relationship between X and Y as X is a self-selected “treatment”.

We recognize that in fMRI research, most experiments for the estimation of ‘causal’ relationships compare a single group of subjects observed under different experimental conditions, not different groups of subjects exposed to different conditions, as here. The framework herein can be extended to this case as well, but to do so would require more space than available in this brief commentary, and is the focus of a separate research article currently in preparation.

The example illustrates that even in a randomized experiment, causal interpretations of RLSEMs (and DGMs) rest on strong untestable assumptions. Some methodological treatments of DGM’s and RLSEMs have recognized this point, but many influential treatments (e.g., Judd and Kenny (1981)) have not. Thus, practitioners often use RLSEMs without realizing the types of assumptions they implicitly make when model coefficients are interpreted as effects. It is also important to point out that other popular methods for studying effective connectivity, for example, DCM (Friston et al (2003)) and Granger causality (Roebroeck et al (2005)), implicitly make similar types of assumptions. The fact that these assumptions go unrecognized is a shortcoming of the field.

We close by discussing alternative assumptions and research designs that allow causal interpretations to be given to RLSEMs and DGMs when applied to neuroimaging data from a randomized experiment. When assumption (4b) is not made, but assumption (2) is made, Sobel (2008) shows how the instrumental variable estimand can be used to estimate the effect of X on Y . Alternatively and preferably, it is sometimes possible to directly manipulate the intermediate outcome. The combination of fMRI and transcranial magnetic stimulation (TMS) promises to integrate the ability of neuroimaging to observe brain activity with the ability of TMS to manipulate brain function (Bohning et al., 1997). Using this technique one can simulate temporary “brain lesions” while the subject performs certain tasks. One can then infer the effect of brain activity on task performance by comparing subjects randomized to receive/not receive stimulation. Ultimately, we believe that techniques like TMS, which allow direct manipulation of intermediate outcomes, provide the most promising approach for studying effective connectivity. That is because such procedures rely on research designs with known properties in lieu of assumptions that are neither testable nor known to be correct.

References

1. Bohning DE, Pecheny AP, Epstein CM, Speer AM, Vincent DJ, Dannels W, George M. Mapping transcranial magnetic stimulation (TMS) fields in vivo with MRI. *Neuroreport*. 1997; 8:2535–2538. [PubMed: 9261822]
2. Friston KJ, Harrison L, Penny W. Dynamic causal modeling. *NeuroImage*. 2003; 19:1273–1302. [PubMed: 12948688]
3. Judd CM, Kenny DA. Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*. 1981; 5:602–619.
4. Neumann J, Fox P, Turner R, Lohmann G. Learning partially directed functional networks from meta-analysis imaging data. *NeuroImage*. 2010; 49:1372–1384. [PubMed: 19815079]
5. Pearl J. Causal diagrams for empirical research (with discussion). *Biometrika*. 1995; 82:669–710.
6. Ramsay, JO.; Silverman, BW. *Functional Data Analysis*. Second Edition. New York: Springer; 2006.

7. Ramsey JD, Spirtes P, Glymour C. On meta-analyses of imaging data and the mixture of records. *NeuroImage*. 2010 this issue.
8. Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods – Application to control of the healthy worker survivor effect. *Mathematical Modeling*. 1986; 7:1393–1512.
9. Robins, JM. Semantics of causal DAG models and the identification of direct and indirect effects. In: Green, P.; Hjort, N.; Richardson, S., editors. *Highly Structured Stochastic Systems*. Oxford University Press; 2003.
10. Roebroeck A, Formisano E, Goebel R. Mapping directed influence over the brain using Granger causality and fMRI. *NeuroImage*. 2005; 25:230–242. [PubMed: 15734358]
11. Sobel ME. Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics*. 2008; 33:230–251.
12. Sobel, ME. Causal Inference in Randomized and Non-Randomized Studies: The Definition, Identification and Estimation of Causal Parameters. In: Milsap, RA.; Maydeu-Olivares, A., editors. *Handbook of Quantitative Methods in Psychology*. Sage; 2009.
13. Spirtes, P.; Glymour, C.; Scheines, R. *Causation, Prediction and Search*. 2nd Edition. MIT Press; 2000.

$$f(x,y,z) = f(y|x)f(x|z)f(z)$$

Figure 1.

A simple three variable graphical model, along with the joint probability distribution it represents.

Table 1

A hypothetical experiment in which subjects 1–3 are treated ($Z = 1$) and subjects 4–6 are not ($Z = 0$). The columns $Y(0)$ and $Y(1)$ contain responses for subjects not receiving and receiving treatment, respectively, and $Y(1) - Y(0)$ is the unit causal effect, which cannot be directly determined from the observed data. Y_{obs} is the observed response.

Subject	Z	$Y(0)$	$Y(1)$	$Y(1) - Y(0)$	Y_{obs}
1	1	-	2.1	-	2.1
2	1	-	1.6	-	1.6
3	1	-	2.0	-	2.0
4	0	0.7	-	-	0.7
5	0	0.5	-	-	0.5
6	0	0.3	-	-	0.3

$$\bar{Y}_{\{Z=1\}} = \frac{2.1+1.6+2.0}{3} = 1.9 \bar{Y}_{\{Z=0\}} = \frac{0.7+0.5+0.3}{3} = 0.5$$