22

Overthrowing the Tyranny of Null Hypotheses Hidden in Causal Diagrams

SANDER GREENLAND

1 Introduction

Graphical models have a long history before and outside of causal modeling. Mathematical graph theory extends back to the 1700s and was used for circuit analysis in the 19th century. Its application in probability and computer science dates back at least to the 1960s (Biggs et al., 1986), and by the 1980s graphical models had become fully developed tools for these fields (e.g., Pearl, 1988; Hajek et al., 1992; Lauritzen, 1996).

As *Bayesian networks*, graphical models are carriers of direct conditional independence judgments, and thus represent a collection of assumptions that confine prior support to a lower dimensional manifold of the space of prior distributions over the nodes. Such dimensionality reduction was recognized as essential in formulating explicit and computable algorithms for digital-machine inference, an essential task of artificial-intelligence (AI) research. By the 1990s, these models had been merged with causal path diagrams long used in observational health and social science (OHSS) (Wright, 1934; Duncan, 1975), resulting in a formal theory of causal diagrams (Spirtes et al., 1993; Pearl, 1995, 2000).

It should be no surprise that some of the most valuable and profound contributions to these *developments* were from Judea Pearl, a renowned AI theorist. He motivated causal diagrams as *causal* Bayesian networks (Pearl, 2000), in which the basis for the dimensionality reduction is grounded in judgments of causal independence (and especially, autonomy) rather than mere probabilistic independence. Beyond his extensive technical and philosophical contributions, Pearl fought steadfastly to roll back prejudice against causal modeling and causal graphs in statistics. Today, only a few statisticians still regard causality as a metaphysical notion to be banned from formal modeling (Lad, 1999). While a larger minority still reject some aspects of causal-diagram or potential-outcome theory (e.g., Dawid, 2000, 2008; Shafer, 2002), the spreading wake of applications display the practical value of these theories, and formal causal diagrams have advanced into applied journals and books (e.g., Greenland et al., 1999; Cole and Hernán, 2002; Hernán et al., 2002; Jewell, 2004; Morgan and Winship, 2007; Glymour and Greenland, 2008) – although their rapid acceptance in OHSS may well have been facilitated by the longstanding informal use of path diagrams to represent qualities of causal systems (e.g., Susser, 1973; Duncan, 1975).

Graphs are unsurpassed tools for illustrating certain mathematical results that hold in functional systems (whether stochastic or not, or causal or not). Nonetheless, it is essential to recognize that many if not most causal judgments in OHSS are based on

observational (purely associational) data, with little or nothing in the way of manipulative (or "surgical") experiment to test these judgments. Time order is usually known, which insures that the chosen arrow directions are correct; but rarely is there a sound basis for deleting an arrow, leaving autonomy in question. When all empirical constraints encoded by the causal network come from passive frequency observations rather than experiments, the primacy of causal independence judgments has to be questioned. In these situations (which characterize observational research), we should not neglect associational models (including graphs) that encode frequency-based judgments, for these models may be all that are identified by available data. Indeed, a deep philosophical commitment to statistically identified quantities seems to drive the arguments of certain critics of potential outcomes and causal diagrams (Dawid, 2000, 2008). Even if we reject this philosophy, however, we should retain the distinction between levels of identification provided by our data, for even experimental data will not identify everything we would like to know.

I will argue that, in some ways, the distinction of nonidentification from identification is as fundamental to modeling and statistical inference about causal effects as is the distinction of causation from association (Gustafson, 2005; Greenland, 2005a, 2009a, 2009b). Indeed, I believe that some of the controversy and confusion over causation versus association stems from the inability of statistical observations to point identify (consistently estimate) many of the causal parameters that astute scientists legitimately ask about. Furthermore, if we consider strategies that force identification from available data (such as node or arrow deletions from graphical models) we will find that identification may arise only by declaring some types of joint frequencies as justifying the corresponding conditional independence assumptions. This leads directly into the complex topic of pruning algorithms, including the choice of target or loss function.

I will outline these problems in their most basic forms, for I think that in the rush to adopt causal diagrams some realism has been lost by neglecting problems of nonidentification and pruning. My exposition will take the form of a series of vignettes that illustrate some basic points of concern. I will not address equally important concerns that many of the nodes offered as "treatments" may be ill-defined or nonmanipulable, or may correspond poorly to the treatments they ostensibly represent (Greenland, 2005b; Hernán, 2005; Cole and Frangakis, 2009; VanderWeele, 2009).

2 Nonidentification from Unfaithfulness in a Randomized Trial

Nonidentification can be seen and has caused controversy in the simplest causal-inference settings. Consider an experiment that randomizes a node R. Inferences on causal effects of R from subsequent associations of R with later events would then be justified, since R would be an exogenous node. R would also be an instrumental variable for certain descendants under further conditional-independence assumptions.

A key problem is how one could justify removing arrows along the line of descent from R to another node Y, even if R is exogenous. The overwhelmingly dominant approach licenses such removal if the observed R-Y association fails to meet some criterion for departure from pure randomness. This schematic for a causal-graph pruning

algorithm was employed by Spirtes et al. (1993), unfortunately with a very naïve Neyman-Pearsonian criterion (basically, allowing removal of arrows when a *P*-value exceeds an α level). These and related graphical algorithms (Pearl and Verma, 1991) produce what appear to be results in conflict with practical intuitions, namely causal "discovery" algorithms for single observational data sets, with no need for experimental evidence. These algorithms have been criticized philosophically on grounds related to the identification problem (Freedman and Humphreys, 1999; Robins and Wasserman, 1999ab), and there are also objections based on statistical theory (Robins et al., 2003).

One controversial assumption in these algorithms is *faithfulness* (or stability) that all connected nodes are associated. Although arguments have been put forward in its favor (e.g., Spirtes et al., 1993; Pearl, 2000, p. 63), this assumption coheres poorly with prior beliefs of some experienced researchers. Without faithfulness, two nodes may be independent even if there is an arrow linking them directly, if that arrow represents the presence of causal effects among units in a target population. A classic example of such unfaithfulness appeared in the debates between Fisher and Neyman in the 1930s, in which they disagreed on how to formulate the causal null hypothesis (Senn, 2004). The framework of their debate would be recognized today as the *potential-outcome* or counterfactual model, although in that era the model (when named) was called the randomization model. This model illustrates the benefit of randomization as a means of detecting a signal by injecting white noise into a system to drown out uncontrolled influences.

To describe the model, suppose we are to study the effect of a treatment X on an outcome Y_{obs} observable on units in a specific target population. Suppose further we can fully randomize X, so X will equal the randomized node R. In the potential-outcome formulation, the outcome becomes a vector **Y** indexed by X. Specifically, X determines which component Y_x of **Y** is observable conditional on X=x: $Y_{obs} = Y_x$ given X=x. To say X can causally affect a unit makes no reference to observation, however; it merely means that some components of **Y** are unequal. With a binary treatment and outcome, there are four types of units in the target population about a binary treatment X which indexes a binary potential-outcome vector **Y** (Copas, 1973):

1) Noncausal units with outcomes **Y**=(1,1) under X=1,0 ("doomed' to Y_{obs}=1);
2) Causal units with outcomes **Y**=(1,0) under X=1,0 (X=1 causes Y_{obs}=1);
3) Causal units with outcomes **Y**=(0,1) under X=1,0 (X=1 prevents Y_{obs}=1); and
4) Noncausal units with outcomes **Y**=(0,0) under X=1,0 ("immune" to Y_{obs}=1).

Suppose the proportion of type *i* in the trial population is p_i. There are now two null hypotheses:

H_s: There are no causal units: $p_2=p_3=0$ (sharp or strong null),

H_w: There is no net effect of treatment on the distribution of Y_{obs}: $p_2=p_3$ (weak null).

Under the randomization distribution we have

$E(Y_{obs}|X=1) = Pr(Y_{obs}=1|do[X=1]) = Pr(Y_1=1) = p_1+p_2$ and

$E(Y_{obs}|X=0) = Pr(Y_{obs}=1|do[X=0]) = Pr(Y_0=1) = p_1+p_3$;

hence H_w: $p_2=p_3$ is equivalent to the hypothesis that the expected outcome is the same for both treatment groups, and that the proportions with Y_{obs}=1 under the extreme population

intervention do[X=1] to every unit and do[X=0] to every unit are equal. Note however that only H_s entails that the proportion with $Y_{obs}=1$ would be the same under *every* possible allocation of treatment X among the units; this property implies that the Y margin is fixed under H_s, and thus provides a direct causal rationale for Fisher's exact test of H_s (Greenland, 1991).

H_s also entails H_w (or, in terms of parameter subspaces, $H_s \subset H_w$). The converse is false; but, under any of the "optimal" statistical tests that can be formulated from data on X and Y_{obs} only, power is identical to the test size on all alternatives to the sharp null with $p_2=p_3$, i.e., H_s is not identifiable within H_w, so within H_w the power of any valid test of H_s will not exceed its nominal alpha level. Thus, following Neyman, it is only relevant to think in terms of H_w, because H_w could be rejected whenever H_s could be rejected. Furthermore, some later authors would disallow $H_w - H_s$: $p_2 = p_3 \neq 0$ because it violates faithfulness (Spirtes et al., 2001) or because it represents an extreme treatment-by-unit interaction with no main effect (Senn, 2004).

There is also a Bayesian argument for focusing exclusively on H_w. H_w is of Lebesgue measure zero, so under the randomization model, distinctions within H_w can be ignored by inferences based on an absolutely continuous prior on $\mathbf{p} = (p_1,p_2,p_3)$ (Spirtes et al., 1993). More generally, any distinction that remains *a posteriori* can be traced to the prior. A more radical stance would dismiss both H_s and the model defined by 1-4 above as "metaphysical," because it invokes constraints on the joint distribution of the components Y_1 and Y_0, and that joint distribution is not identified by randomization of X if only X and Y_{obs} are observed (Dawid, 2000).

On the other hand, following Fisher one can argue that the null of key scientific and practical interest is H_s, and that $H_w - H_s$: $p_2 = p_3 \neq 0$ is a scientifically important and distinct hypothesis. For instance, $p_2>0$, $p_3>0$ entails the existence of units who should be treated quite differently, and provides an imperative to seek covariates that discriminate between the two causal types, even if $p_2=p_3$. Furthermore, rejection of the stronger H_s is a *weaker* inference than rejection of the weaker H_w, and thus rejecting only H_s would be a conservative interpretation of a "significant" test statistic. Thus, focusing on H_s is compatible with a strictly falsificationist view of testing in which acceptance of the null is disallowed. Finally, there are real examples in which X=1 causes Y=1 in some units and causes Y=0 in others; in some of these cases there may be near-perfect balance of causation and prevention, as predicted by certain physical explanations for the observations (e.g., as in Neutra et al., 1980).

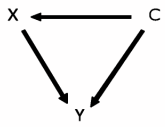
To summarize, identification problems arose in the earliest days of formal causal modeling, even when considering only the simplest of trials. Those problems pivoted not on whether one should attempt formal modeling of causation as distinct from association, but rather on what could be identified by standard experimental designs. In the face of limited (and limiting) design strategies, these problems initiated a long history of attempts to banish identification problems based on idealized inference systems and absolute philosophical assertions. But a counter-tradition of arguments, both practical and philosophical, has regarded identification problems as carriers of valuable scientific information: They are signs of study limitations which need to be recognized and can

only be dealt with effectively by innovative data collection (e.g., measuring more covariates or deploying new study designs), instead of by increasing sample sizes and defining the problems away so that "identical replications" are sufficient to narrow inferences.

3 Causal Diagrams Encode Numerous Uncertain Null Hypotheses

To move to the observational setting that is my main concern, consider figure 1, a typical causal diagram used to illustrate assumptions used by methods for estimating "the effect of X on Y" from observational data.

Figure 1: Naïve causal diagram



The first point to note is that this diagram is woefully incomplete relative to the epidemiologic reality, because it ignores

a) unmodeled confounders (variables not in the graph that affect more than one node in the graph),
b) selection effects (effects of factors in the graph on selection), and
c) measurement errors (which require addition of measurement nodes for each imperfectly measured node).

Put another way, typical causal DAGs like that in figure 1 are full of hidden, assumed null hypotheses, in the form of assumptions that imply problems a, b, and c are absent. For example, a causal DAG assumes that for **every** node pair (A,B) in the DAG,

1) there is **no** shared ancestor not in graph (not A↔B),
2) there is **no** unmarked conditioning event that has opened a path between A and B (not A─B),
3) if A and B are nonadjacent (neither A→B nor A←B), there is **no** mechanism that leads directly from one node to another (thus bypassing other nodes in the graph).

Not every study will seriously violate all of these assumptions. But in most studies in OHSS, none of the nulls 1-3 will have convincing support, and any purported test of a causal effect will really be a test of these 3 nulls as well as the specified causal null. This fact is just a special case of longstanding observations that statistical tests are really tests of all assumptions used in the test, not just the particular null of interest (Fisher, 1943; Box, 1980). In this regard, note that absence of arrows between nodes (3) encodes particularly strong nulls that are routinely presumed but rarely have supporting data. More often in OHSS, we observe only a conditional temporal sequence such as "A precedes B," which may be due to A→B, A↔B, A─B or some combination.

While sensitivity analysis is often recommended to examine the impact of deviations from assumptions, it becomes unintelligible if not infeasible as the number of assumptions (or corresponding parameters) increase. Then too, some causal inferences will display unlimited sensitivity to certain assumptions, requiring the introduction of priors on the corresponding parameters in order to salvage any inference (Greenland, 1998, 2005a; Gustafson, 2005). This problem arises in the model given below.

4 Eliminating Unsupported Nulls (graphical realism)

Let conditioning be denoted with square brackets around the conditioned event node. Then, in contrast to Figure 1, realistic causal graphs for OHSS will have

1) numerous unobserved (latent) nodes, often more of them than observed nodes,
2) few node pairs without an arc between them,
3) no **observed** set of variables sufficient for bias control, and
4) a selection node S that is bracketed and potentially affected by most other nodes.

In particular, when all variables are subject to measurement error, a realistic causal model for a single exposure-disease analysis will have at least:

X = Exposure, X*: measured X

Y = Outcome, Y*: measured Y

C = Known antecedents, C*: measured C

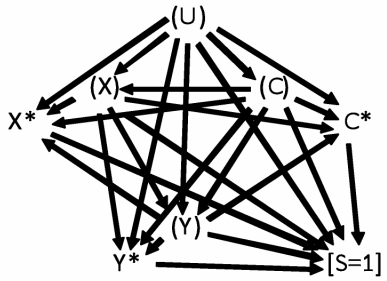
U = Other antecedents (unmeasured and possibly unknown)

S = Selection into the analysis (from selection into the study plus exclusions).

Because analysis is always conditioned on S=1, we should always show this conditioning event on the graph with a circle or brackets around it, e.g., as [S=1].

As an example, fig. 2 shows what I'd consider a **minimal** realistic causal graph for a typical case-control study of a life history and a degenerative disease outcome (e.g, nicotine intake X and Alzheimer's disease Y), which has 25 of 28 possible adjacencies.

Figure 2: Realistic causal diagram



What can fig. 2 provide if further assumptions cannot be justified? The only observed distribution is p(c*,x*,y*|S=1), which is not a factor in the causal Markov decomposition entailed by the graph,

$p(u,c,x,y,c^*,x^*,y^*,s) =$

$p(u)p(c|u)p(x|u,c)p(y|u,c,x)p(c^*|u,c,x,y)p(x^*|u,c,x,y)p(y^*|u,c,x,y)p(s|u,c,x,y,c^*,x^*,y^*)$,

which involves both S=0 events (not selected) and S=1 events (selected), i.e., the lowercase "s" is used when S can be either 0 or 1.

The marginal (total-population) potential-outcome distribution for Y after intervention on X, $p(y_x)$, equals $p(y|do[X=x])$, which under fig. 2 equals the standardized (mixing) distribution of Y given X standardized to (weighted by or mixed over) $p(c,u) = p(c|u)p(u)$:

$p(y_x) = p(y|do[x]) = \sum_{u,c} p(y|u,c,x)p(c|u)p(u)$.

This estimand involves only three factors in the decomposition, but none of them are identified if U is unobserved and no further assumptions are made. Analysis of the causal estimand $p(y_x)$ must somehow relate it to the observed distribution $p(c^*,x^*,y^*|S=1)$ using known or estimable quantities, or else remain purely speculative (i.e., a sensitivity analysis).

It is a long, hard road from $p(c^*,x^*,y^*|S=1)$ to $p(y_x)$, much longer than the current "causal inference" literature often makes it look. To appreciate the distance, rewrite the summand of the standardization formula for $p(y_x)$ as an inverse-probability-weighted (IPW) term derived from an observation $(c^*,x^*,y^*|S=1)$: From fig. 2,

$p(y|u,c,x)p(c|u)p(u) =$

$p(c^*,x^*,y^*|S=1)p(S=1)p(u,c,x,y|c^*,x^*,y^*,S=1)/$

$p(x|u,c)p(c^*|u,c,x,y)p(x^*|u,c,x,y)p(y^*|u,c,x,y)p(S=1|u,c,x,y,c^*,x^*,y^*)$.

The latter expression includes

1) the exposure dependence on its parents, $p(x|u,c)$;
2) the measurement distributions $p(c^*|u,c,x,y)$, $p(x^*|u,c,x,y)$, $p(y^*|u,c,x,y)$; and
3) the fully conditioned selection probability $p(S=1|u,c,x,y,c^*,x^*,y^*)$.

The absence of effects corresponding to 1−3 from graphs offered as "causal" suggests that "causal inference" from observational data using formal causal models remains a theoretical and largely speculative exercise (albeit often presented without explicit acknowledgement of that fact).

When adjustments for these effects are attempted, we are usually forced to use crude empirical counterparts of terms like those in 1−3, with each substitution demanding nonidentified assumptions. Consider that, for valid inference under figure 2,

1) Propensity scoring and IPW for treatment need $p(x|u,c)$, but all we get from data is $p(x^*|c^*)$. Absence of u and c is usually glossed over by assuming "no unmeasured confounders" or "no residual confounding." These are not credible assumptions in OHSS.
2) IPW for selection and censoring needs $p(S=1|u,c,x,y,c^*,x^*,y^*)$, but usually the most we get from a cohort study or nested study is $p(S=1|c^*,x^*)$. We do not even get that much in a case-control study.
3) Measurement-error correction needs conditional distributions from $p(c^*,x^*,y^*,u,c,x,y|S=1)$, but even when a "validation" study is done, we obtain only alternative measurements $c^\dagger,x^\dagger,y^\dagger$ (which are rarely error-free) on a tiny and

biased subset. So we end up with observations from $p(c^{\dagger},x^{\dagger},y^{\dagger},c^*,x^*,y^*|S=1,V=1)$ where V is the validation indicator.

4) Consistency between the observed X and the intervention variable, in the sense t hat $P(Y|X=x) = P(Y|do[X=x],X=x)$. This can be hard to believe for common variables such as smoking, body-mass index, and blood pressure, even if do[X=x] is well-defined (which is not usually the case).

In the face of these realities, standard practice seems to be: Present wildly hypothetical analyses that pretend the observed distribution $p(c^*,x^*,y^*|S=1)$, perhaps along with $p(c^{\dagger},x^{\dagger},y^{\dagger},c^*,x^*,y^*|S=1,V=1)$ or $p(S=1|c^*,x^*)$, is sufficient for causal inference. The massive gaps are filled in with models or assumptions, which are priors that reduce dimensionality of the problem to something within computing bounds. For example, use of IPW with $p(S=1|c^*,x^*)$ to adjust for selection bias (as when 1−S is a censoring indicator) depends crucially on a nonidentified ignorability assumption that $S \perp\!\!\!\perp (U,C,X,Y)|(C^*,X^*)$, i.e., that selection S is independent of the latent variables U,C,X,Y given the observed variables C*,X*. We should expect this condition to be violated whenever a latent variable affects selection directly or shares unobserved causes with selection. If such effects are exist but are missing from the analysis graph, then by some definitions the graph (and hence the resulting analysis) isn't causal, no matter how much propensity scoring (PS), marginal structural modeling (MSM), inverse-probability weighting (IPW), or other causal-modeling procedures we apply to the observations $(c^*,x^*,y^*|S=1)$.

Of course, the overwhelming dimensionality of typical OHSS problems virtually guarantees that arbitrary constraints will enter at some point, and forces even the best scientists to rely on a tiny subset of all the models or explanations consistent with available facts. Personal bias in determining this subset may be unavoidable due to strong cultural influences (such as adherence to received theories, as well as moral strictures and financial incentives), which can also lead to biased censoring of observations (Greenland, 2009c). One means of coping with such bias is by being aware of it, then trying to test it against the facts one can muster (which are often few).

The remaining sections sketch some alternatives to pretending we can identify unbiased or assuredly valid estimators of causal effects in observational data, as opposed to within hypothetical models for data generation (Greenland, 1990; Robins, 2001). In these approaches, both frequentist and Bayesian analyses are viewed as hypotheticals conditioned on a data-generation model of unknown validity. Frequentist analysis provides only inferences of the form "if the data-generation process behaves like this, here is how the proposed decision rule would perform," while Bayesian analysis provides only inferences of the form "if I knew that its data-generation process behaves like this, here is how this study would alter my bets."[1] If we aren't sure how the data-generation

[1]This statement describes Bayes factors (Good, 1983) conditioned on the model. That model may include an unknown parameter that indexes a finite number of submodels scattered over some high-dimensional subspace, in which case the Bayesian analysis is called "model averaging," usually with an implicit uniform prior over the models. Model averaging may also operate over continuous parameters via priors on those parameters.

process behaves, no statistical analysis can provide more, no matter how much causal modeling is done.

5 Predictive Analysis

If current models for observed-data generators (whether logistic, structural, or propensity models) can't be taken seriously as "causal", what can we make of their outputs? It is hard to believe the usual excuses offered for regression outputs (e.g., that they are "descriptive") when the fitted model is asserted to be causal or "structural." Are we to consider the outputs of (say) and IPW-fitted MSM to be some sort of data summary? Or will it function as some kind of optimal predictor of outcomes in a purely predictive context? No serious case has been made for causal models in either role, and it seems that some important technical improvements are needed before causal modeling methods become credible predictive tools.

Nonetheless, graphical models remain useful (and might be less misleading) even when they are not "causal," serving instead as mere carriers of conditional independence assumptions within a time-ordered framework. In this usage, one may still employ presumed causal independencies as prior judgments for specification. In particular, for predictive purposes, some or all of the arrows in the graph may retain informal causal interpretations; but they may be causally wrong, and yet the graph can still be correct for predictive purposes.

In this regard, most of the graphical modeling literature in statistics imposes little in the way of causal burden on the graph, as when graphs are used as influence diagrams, belief and information networks, and so on without formal causal interpretation (that is, without representing a formal causal model, e.g., Pearl, 1988; Hajek et al., 1992; Cox and Wermuth, 1996; Lauritzen, 1996). DAG rules remain valid for prediction if the absence of an open path from X to Y is interpreted as entailing $X \perp\!\!\!\perp Y$, or equivalently if the absence of a directed path from X to Y (in causal terms, X is not a cause of Y; equivalently, Y is not affected by X) is interpreted as entailing $X \perp\!\!\!\perp Y|\mathbf{pa}_X$, the noncausal Markov condition (where $\mathbf{pa}_X$ is the set of parents of X). In that case, $X \rightarrow Y$ can be used in the graph even if X has no effect on Y, or vice-versa.

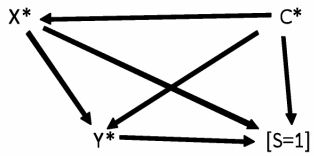
As an example, suppose X and Y are never observed without them affecting selection S, as when X is affects miscarriage S and Y is congenital malformation. If the target population is births, X predicts malformations Y among births (which have S=1). As another example, suppose X and Y are never observed without an uncontrolled, ungraphed confounder U, as when X is diet and Y is health status. If one wishes to target those at high risk for screening or actuarial purposes it does not matter if $X \rightarrow Y$ represents a causally confounded relation. Lack of a directed path from X to Y now corresponds to lack of additional predictive value for Y from X given $\mathbf{pa}_X$. Arrow directions in temporal (time-ordered) predictive graphs correspond to point priors about time order, just as they do in causal graphs.

Of course, if misinterpreted as causal, predictive inferences from graphs (or any predictive modeling) may be potentially disastrous for judging interventions on X. But, in OHSS, the causality represented by a directed path in a so-called causal diagram rarely

corresponds to more than a hypothesis, plausible perhaps but only one among a myriad of others. If most arrows shown in a graph encode no real data other than an observed conditional temporal sequencing, then labeling the graph as a "causal diagram" sets the stage for the disaster.

Figure 3 is the temporal predictive diagram for the observables in the earlier example, assuming those events occur in the order C*, X*, Y*, [S=1].

Figure 3: Temporally predictive diagram



Comparison to the causal diagram in figure 2 illustrates how a temporal predictive diagram for an observable frequency distribution may be derived from an underlying causal diagram for a nonidentified theory. Figure 3 is saturated in the sense that all nodes are connected by an edge, but this need not be so for a predictive diagram derived from a causal one. If there is temporal ambiguity among the observables, there may be multiple predictive diagrams compatible with the causal diagram (which will form a subset of the multiple probability graphs compatible with the causal diagram).

If we treat causal models as carriers of prior information about conditional independencies, they appear as legitimate candidates to consider as predictive models. For example, MSMs can be evaluated as devices for prediction from fixed sequences and structural nested models can be evaluated as devices for prediction from stochastic processes. I would thus offer this challenge to the current "longitudinal causal modeling" literature: If we know our observations are just a dim and distant projection of the causal structure and we can only identify predictive links among observed quantities, are there predictive advantages of structural modeling (modeling potential outcomes as well as observed outcomes)? If not, what precisely is the advantage of fitting such models (compared to noncausal models) when effects are not identified?

I believe there *are* advantages of causal models, precisely as described by Pearl (2000): They provide an encoding for qualitative (structural) prior information expressed in terms of "cause" and "effect." But in current practice, fitting methods for complex causal models are quite primitive, and need to incorporate properly smoothness and other information that can be freely assumed in purely predictive-modeling approaches. This is a general problem of semi-parametric theory: It necessarily focuses sharp constraints in some dimensions and none in "most" dimensions (represented by the infinite-dimensional time component in standard Cox models). When relevant dimensions for constraint (those

where much background information is available) are not well represented by the dimensions constrained by the model, considerably efficiency can be lost for estimating parameters of interest. A striking example given by Whittemore and Keller (1986) displayed the poor small-sample performance for estimating a survival curve when using an unsmoothed nonparametric hazard estimator (Kaplan-Meier or Nelson-Altschuler estimation), relative to spline smoothing of the hazard.

6 Pruning the Identified Portion of the Model

Over recent decades, great strides have been made in creating predictive algorithms; the question remains however, what role should these algorithms play in causal inference? It would seem that these algorithms can be beneficially applied to fitting the marginal distribution identified by the observations. Nonetheless, the targets of causal inference in observational studies lie beyond the identified margin, and thus beyond the reach of these algorithms. At best, then, the algorithms can provide the identified foundation for building into unobserved dimensions of the phenomena under study.

Even if we focus only on the identified margin, however, there may be far more nodes and edges than seem practical to allow in the final model. A prominent feature of modern predictive algorithms is that they start with an impractically large number of terms and then aggressively prune the model, and may re-grow and re-prune repeatedly (Hastie et al., 2009). These strategies coincide with the intuition that omitting a term is justified when its contribution is too small to stand out against bias and background noise; e.g., we do not include variables like patient identification number because we know that are usually pure noise.

Nonetheless, automated algorithms often delete variables or connections that prior information instead suggests are relevant or related; thus shields from pruning are often warranted. Furthermore, a deleted node or arrow may indeed be important from a contextual perspective even if does not meet algorithmic retention criteria. Thus, model simplification strategies such as pruning may be justified by a need for dimensionality reduction, but should be recognized as part of algorithmic compression or computational prediction, not as a mode of inference about structural models.

Apart from these vague cautions, it has long been recognized that if our goal is to evaluate causal effects, different loss functions are needed from those in the pruning algorithms commonly applied by researchers. Specifically, the loss or benefit entailed by pruning needs to be evaluated in reference to the target effect under study, and not simply successful prediction of identified quantities. Operationalizing this imperative requires building out into the nonidentified (latent) realm of the target effects, which is the focus of *bias modeling*.

7 Modeling Latent Causal Structures (Bias Modeling)

The target effects in causal inference are functions of unobserved dimensions of the data-generating process, which consist primarily of bias sources (Greenland, 2005a). Once we recognize the nonidentification this structure entails, the major analysis task shifts away

from mathematical statistics to prior specification, because with nonidentification only proper priors on nonidentified parameters can lead to proper posteriors.

Even the simplest point-exposure case can involve complexities that transform simple and precise-looking conventional results into complex and utterly ambiguous posteriors (Greenland, 2009a, 2009b). In a model complex enough to reflect Figure 2, there are far too many elements of specification to contextually justify them all in detail. For example, one could only rarely justify fewer than two free structural parameters per arrow, and the distributional form for each parameter prior would call for at least two hyperparameters per parameter (e.g., a mean and a variance), leading to at least 50 parameters and 100 hyperparameters in a graph with 25 arrows. Allowing but one prior association parameter (e.g., a correlation) per parameter pair adds over 1,000 (50 choose 2) more hyperparameters.

As a consequence of the exponential complexity of realistic models, prior specification is difficult, ugly, ad hoc, highly subjective, and tentative in the extreme. In addition, the hard-won model will lack generalizability and elegance, making it distasteful to both the applied scientist and the theoretical statistician. Nor will it please the applied statistician concerned with "data analysis," for the analysis will instead revolve around numerous contextual judgments that enlist diverse external sources of information. In contrast to the experimental setting (in which the data-generation model may be dictated entirely by the design), the usually sharp distinction between prior and data information will be blurred by the dependence of the data-generation model on external information.

These facts raise another challenge to the current "causal modeling" literature: If we know our observations are just a dim and distant projection of the causal structure and we can only identify predictive links among observed quantities, how can we incorporate simultaneously all error sources (systematic as well as random) known to be important into a complex longitudinal framework involving mismeasurement of entire sequences of exposures and confounders? Some progress on this front has been made, but primarily in contexts with validation data available (Cole et al., 2010), which is not the usual case.

8 The Descriptive Alternative

In the face of the extraordinary complexity of realistic models for OHSS, it should be an option of each study to focus on describing the study and its data thoroughly, sparing us attempts at inference about nonidentified quantities such as "causal effects." This option will likely never be popular, but should be allowed and even encouraged (Greenland et al., 2004). After all, why should I care about your causal inferences, especially if they are based on or grossly over-weighted by the one or few studies that you happened to be involved in? If I am interested in forming my own inferences, I do want to see your data and get an accurate narrative of the physical processes that produced them. In this regard, statistics may supply data summaries. Nonetheless, it must be made clear exactly how the statistics offered reflect the data as opposed to some hypothesis about the population from which they came; *P*-values do not satisfy this requirement (Greenland, 1993; Poole, 2001).

Here then is a final challenge to the "causal modeling" literature: If we know our observations are just a dim and distant projection of the causal structure and we can only identify associations among observed quantities, how can we interpret the outputs of "structural modeling" (such as confidence limits for ostensibly causal estimands which are not in fact identified) as data summaries? We should want to see answers that are sensible when the targets are effects in a context at least as complex as in fig. 2.

9 What is a Causal Diagram?

The above considerations call into question some epidemiological accounts of causal diagrams. Pearl (2000) describes a causal model *M* as a formal functional system giving relations among a set of variables. *M* defines a joint probability distribution p() and an intervention operator do[] on the variables. A causal diagram is then a directed graph *G* that implies the usual Markov decomposition for p() and displays additional properties relating p() and do[]. In particular, each child-parent family {X, $\mathbf{pa}_X$} in *G* satisfies

1) p(x|do[$\mathbf{pa}_X$=a]) = p(x|$\mathbf{pa}_X$=a), and
2) if Z is not in {X, $\mathbf{pa}_X$}, p(x|do[Z=z],$\mathbf{pa}_X$=a) = p(x|$\mathbf{pa}_X$=a).

(e.g., see Pearl, 2000, p. 24). These properties stake out *G* as an illustration (mapping) of structure within *M*.

Condition 1 is often described as stating that the association of each node X with its parent vector $\mathbf{pa}_X$ is unconfounded given *M*. Condition 2 says that, given *M*, the only variables in *G* that affect a node X are its parents, and is often called the causal Markov condition (CMC). Nonetheless, as seems to happen often as time passes and methods become widely adopted, details have gotten lost. In the more applied literature, causal diagrams have come to be described as "unconfounded graphs" without reference to an underlying causal model (e.g., Hernán et al., 2004; VanderWeele and Robins, 2007; Glymour and Greenland, 2008). This description not only misses the CMC (2) but, taken literally, means that all shared causes are in the graph.

Condition 1 is a property relating two mathematical objects, *G* and *M*. To claim a diagram is unconfounded is to instead make a claim about the relation of *G* the real world, thus inviting confusion between a *model* for causal processes and the actual processes. For many experts in OHSS, the claim of unconfoundedness has zero probability of being correct because of its highly restrictive empirical content (e.g., see Robins and Wasserman, 1999ab). At best, we can only hope that the diagram provides a useful computing aid for predicting the outcomes of intervention strategies.

As with regression models, causal models in OHSS are always false. Because we can never know we have a correct model (and in fact in OHSS we can't even know if we are very close), to say *G* is causal if unconfounded is a scientifically vacuous definition: It is saying the graph is causal if the causal model it represents is correct. This is akin to saying a monotone increasing function from the range of X to [0,1] is not a probability distribution if it is not in fact how X is distributed; thus a normal(μ,σ^2) cumulative function wouldn't be a probability distribution unless it is *the* actual probability distribution for X (whether that distribution is an objective event generator or a subjective betting schedule).

So, to repeat: To describe a causal diagram as an "unconfounded graph" blurs the distinction between models and reality. Model-based deductions are logical conditionals of the form "model *M* deductively yields these conclusions," and have complete certainty *given* the model *M*. But the model, and hence reality, is never known with certainty, and in OHSS cannot be claimed as known except in the most crude fashion. The point is brought home above by appreciating just how unrealistic all causal models and diagrams in OHSS must be. Thus I would encourage the description of causal diagrams as graphical causal models (or more precisely, graphical representations of certain equivalence classes of causal models), rather than as "unconfounded graphs" (or similar phrases). This usage might even be acceptable to some critics of the current causal-modeling literature (Dawid, 2008).

10 Summary and Conclusions

I would be among the last to deny the utility of causal diagrams; but I argue that their practical utility in OHSS is limited to (i) compact and visually immediate representation of assumptions, and (ii) illustration of sources of nonidentification and bias given realistic assumptions. Converse claims about their utility for identification seem only the latest in a long line of promises to "solve" the problem of causal inference. These promises are akin to claims of preventing and curing all cancers; while progress is possible, the enormous complexity of real systems should leave us skeptical about claims of "solutions" to the real problem.

Many authors have recognized that the problem of effect identification is unsolvable in principle. Although this logical impossibility led some to deny the scientific merit of causal thinking, it has not prevented development of useful tools that have causal-modeling components. Nonetheless, the most precision we can realistically hope for estimating effects in OHSS is about one-digit accuracy, and in many problems even that seems too optimistic. Thus some practical sense is needed to determine what is and isn't important to include as model components. Yet, despite the crudeness of OHSS, good sense seems to lead almost inevitably to including more components than can be identified by available data.

My main point is that effect identification (in the frequentist sense of identification by the observed data) should be abandoned as a primary goal in causal modeling in OHSS. My reasons are practical: Identification will often demand dropping too much of importance from the model, thus imposing null hypotheses that have no justification in either past frequency observations or in priors about mechanisms generating the observations, thus leading to overconfident and biased inferences. In particular, defining a graph as "causal" if it is unconfounded assumes a possibly large set of causal null hypotheses (at least two for every pair of nodes in the graph: no shared causes or conditioned descendants not in the graph). In OHSS, the only graphs that satisfy such a definition will need many latent nodes to be "causal" in this sense, and as a consequence will reveal the nonidentified nature of target effects. Inference may then proceed by imposing contextually defensible priors or penalties (Greenland, 2005a, 2009a, 2009b, 2010).

Despite my view and similar ones (e.g., Gustafson, 2005), I suspect the bulk of causal-inference statistics will trundle on relying exclusively on artificially identified models. It will thus be particularly important to remember that just because a method is labeled a "causal modeling" method does not mean it gives us estimates and tests of actual causal effects. For those who find identification too hard to abandon in formal analysis, the only honest recourse is to separate identified and nonidentified components of the model, focus technique on the identified portion, and leave the latent residual as a topic for sensitivity analysis, speculative modeling, and further study. In this task, graphs can be used without the burden of causality if we allow them a role as pure prediction tools, and they can also be used as causal diagrams of the largely latent structure that generates the data.

Acknowledgments: I am most grateful to Tyler VanderWeele, Jay Kaufman, and Onyebuchi Arah for their extensive and useful comments on this chapter.

References

Biggs, N., Lloyd, E. and Wilson, R. (1986). *Graph Theory, 1736-1936*. Oxford University Press.

Box, G.E.P. (1980). Sampling and Bayes inference in scientific modeling and robustness. *Journal of the Royal Statistical Society*, *Series A* **143**, 383–430.

Cole S.R. and M.A. Hernán (2002). Fallibility in estimating direct effects. *International Journal of Epidemiology* **31**, 163–165.

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

Cole, S.R., L.P. Jacobson, P.C. Tien, L. Kingsley, J.S. Chmiel and K. Anastos (2010). Using marginal structural measurement-error models to estimate the long-term effect of antiretroviral therapy on incident AIDS or death. *American Journal of Epidemiology* **171**, 113-122.

Copas, J.G. (1973). Randomization models for matched and unmatched 2x2 tables. *Biometrika* **60**, 267-276.

Cox, D.R. and N. Wermuth. (1996). *Multivariate Dependencies: Models, Analysis and Interpretation*. Boca Raton, FL: CRC/Chapman and Hall.

Dawid, A.P. (2000). Causal inference without counterfactuals (with discussion). *Journal of the American Statistical Association* **95**, 407-448.

Dawid, A.P. (2008). Beware of the DAG! In: *NIPS 2008 Workshop Causality: Objectives and Assessment*. JMLR Workshop and Conference Proceedings.

Duncan, O.D. (1975). *Introduction to Structural Equation Models*. New York: Academic Press.

Fisher, R.A. (1943; reprinted 2003). Note on Dr. Berkson's criticism of tests of significance. *Journal of the American Statistical Association* **38**, 103–104. Reprinted in the *International Journal of Epidemiology* **32**, 692.

Freedman, D.A. and Humphreys, P. (1999). Are there algorithms that discover causal structure? *Synthese* **121**, 29–54.

Glymour, M.M. and S. Greenland (2008). Causal diagrams. Ch. 12 in: Rothman, K.J., S. Greenland and T.L. Lash, eds. *Modern Epidemiology*, 3rd ed. Philadelphia: Lippincott.

Good, I.J. (1983). *Good thinking*. Minneapolis: U. Minnesota Press.

Greenland, S. (1990). Randomization, statistics, and causal inference. *Epidemiology* **1**, 421-429.

Greenland, S. (1991). On the logical justification of conditional tests for two-by-two-contingency tables. *The American Statistician* **45**, 248-251.

Greenland, S. (1993). Summarization, smoothing, and inference. *Scandinavian Journal of Social Medicine* **21**, 227-232.

Greenland, S. (1998). The sensitivity of a sensitivity analysis. In: 1997 Proceedings of the Biometrics Section. Alexandria, VA: American Statistical Association, 19-21.

Greenland, S. (2005a). Epidemiologic measures and policy formulation: Lessons from potential outcomes (with discussion). *Emerging Themes in Epidemiology* (online journal) 2:1–4. (Originally published as "Causality theory for policy uses of epidemiologic measures," Chapter 6.2 in: Murray, C.J.L., J.A. Salomon, C.D. Mathers and A.D. Lopez, eds. (2002) *Summary Measures of Population Health*. Cambridge, MA: Harvard University Press/WHO, 291-302.)

Greenland, S. (2005b). Multiple-bias modeling for analysis of observational data (with discussion). *Journal of the Royal Statistical Society, Series A* **168**, 267–308.

Greenland, S. (2009a). Bayesian perspectives for epidemiologic research. III. Bias analysis via missing-data methods. *International Journal of Epidemiology* **38**, 1662–1673.

Greenland, S. (2009b). Relaxation penalties and priors for plausible modeling of nonidentified bias sources. *Statistical Science* **24**, 195-210.

Greenland, S. (2009c). Dealing with uncertainty about investigator bias: disclosure is informative. *Journal of Epidemiology and Community Health* **63**, 593-598.

Greenland, S. (2010). The need for syncretism in applied statistics (comment on "The future of indirect evidence" by Bradley Efron). *Statistical Science* **25**, in press.

Greenland, S., J. Pearl, and J.M. Robins (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10**, 37-48.

Greenland, S., M. Gago-Dominguez, and J.E. Castellao (2004). The value of risk-factor ("black-box") epidemiology (with discussion). *Epidemiology* **15**, 519-535.

Gustafson, P. (2005). On model expansion, model contraction, identifiability, and prior information: two illustrative scenarios involving mismeasured variables (with discussion). *Statistical Science* **20**, 111-140.

Hajek, P., T. Havranek and R. Jirousek (1992). *Uncertain Information Processing in Expert Systems*. Boca Raton, FL: CRC Press.

Hastie, T., R. Tibshirani and J. Friedman (2009). *The elements of statistical learning: Data mining, inference, and prediction*, 2nd ed. New York: Springer.

Hernán, M.A. (2005). Hypothetical interventions to define causal effects—afterthought or prerequisite? *American Journal of Epidemiology* **162**, 618–620.

Hernán M.A., S. Hernandez-Diaz, M.M. Werler and A.A. Mitchell. (2002). Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *American Journal of Epidemiology* **155**, 176–184.

Hernán M.A., S. Hernandez-Diaz and J.M. Robins (2004). A structural approach to selection bias. *Epidemiology* **15**, 615-625.

Jewell, N. (2004). *Statistics for Epidemiology*. Boca Raton, FL: Chapman and Hall/CRC.

Lad, F. (1999). Assessing the foundations of Bayesian networks: A challenge to the principles and the practice. *Soft Computing* **3**, 174-180.

Lauritzen, S. (1996). *Graphical Models*. Oxford: Clarendon Press.

Leamer, E.E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

Morgan, S.L. and C. Winship. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.

Neutra, R.R., S. Greenland, and E.A. Friedman (1980). The effect of fetal monitoring on cesarean section rates. *Obstetrics and Gynecology* **55**, 175-180.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

Pearl, J. (1995). Causal diagrams for empirical research (with discussion). *Biometrika* **82**, 669-710.

Pearl, J. (2000; 2nd ed. 2009). *Causality*. New York: Cambridge University Press.

Pearl, J. and P. Verma (1991). A theory of inferred causation. In: *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, Ed. J.A. Allen, R. Filkes and E. Sandewall. San Francisco: Morgan Kaufmann, 441-452.

Poole, C. (2001). Poole C. Low P-values or narrow confidence intervals: Which are more durable? *Epidemiology* **12**, 291–294.

Robins, J.M. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology* **12**, 313–320.

Robins, J.M. and L. Wasserman (1999a). On the impossibility of inferring causation from association without background knowledge. In: *Computation, Causation, and Discovery*. Glymour, C. and Cooper, G., eds. Menlo Park, CA, Cambridge, MA: AAAI Press/The MIT Press, pp. 305-321.

Robins, J.M. and L. Wasserman (1999b). Rejoinder to Glymour and Spirtes. In: *Computation, Causation, and Discovery*. Glymour, C. and Cooper, G., eds. Menlo Park, CA, Cambridge, MA: AAAI Press/The MIT Press, pp. 333-342.

Robins, J.M., R. Scheines, P. Spirtes and L. Wasserman (2003). Uniform consistency in causal inference. *Biometrika* **90**, 491-515.

Senn, S. (2004). Controversies concerning randomization and additivity in clinical trials. *Statistics in Medicine* **23**, 3729–3753.

Shafer, G. (2002). Comment on "Estimating causal effects," by George Maldonado and Sander Greenland. *International Journal of Epidemiology* **31**, 434-435.

Spirtes, P., C. Glymour and R. Scheines (1993; 2nd ed. 2001). *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.

Susser, M. (1973). *Causal Thinking in the Health Sciences*. New York: Oxford University Press.

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

VanderWeele, T.J. and J.M. Robins (2007). Directed acyclic graphs, sufficient causes and the properties of conditioning on a common effect. *American Journal of Epidemiology* **166**, 1096-1104.

Whittemore, A.S. and J.B. Keller (1986). Survival estimation using splines. *Biometrics* **42**, 495-506.

Wright, S., (1934). The method of path coefficients. *Annals of Mathematical Statistics* **5**,161-215.