Causal Inference as a Prediction Problem:

Assumptions, Identification, and Evidence Synthesis

(with corrections to print version p. 13)

Sander Greenland

Department of Epidemiology and

Department of Statistics

University of California, Los Angeles, CA 90095-1772, USA

lesdomes@ucla.edu

Ch. 5 in: Berzuini, C., Dawid, A.P., and Bernardinelli, L. (eds.).

Causal Inference: Statistical Perspectives and Applications. New York: Wiley, 2012, 43-58.

Introduction

I thank the editors for their kind and persistent invitation to contribute an essay. What follows is not a scholarly overview (in the sense of laying out history and development with accuracy and detailed citation), and provides no formal development. Instead it is only a brief, informal outline of my perspective on statistical treatments of causal inference that have emerged over the past several decades, and discussion of problems that seem important in light of my applied experience. That experiences and hence my comments are limited to health and medical research. Those who dislike philosophical treatments (which allow poorly constrained speculations, belaboring the obvious, and obvious errors for the sake of portraying analogies between models, language, and reality) as opposed to mathematical treatments (which allow only sound deduction within axiomatically constrained systems) are advised to read no further. Those who do read on are recommended to Greenland (2010) for more elaboration of several points herein.

I especially emphasize that in applications, I have often found formal developments in causal inference could play only minor if valuable roles. Among the reasons for this limitation are that the simplifications required to make the formal machinery work are often too extreme to yield credible results (Greenland, 2010). Furthermore, currently accepted machinery does not effectively incorporate the highly disparate evidence streams that have to be considered in risk assessment, except as externally imposed model constraints or prior distributions. While there have been many methodologies proposed for combining disparate evidence sources, as yet none have seen widespread adoption. I hope that (despite its incoherencies) my discussion and anecdotes will suggest reasons for that state of affairs, both in terms of the complexity of the problem and human limitations in facing that complexity. Throughout, I will assume readers are familiar with basic causal models.

A Brief Commentary on Developments Since 1970

It is remarkable to look back on the four decades I have been in health-science research and see the transformation of its theoretical foundation for causal inference. By the 1970s key formal elements in that foundation (chiefly potential outcomes, causal diagrams, structural equations) had been around in some form for a half century and seen use in the social sciences. Yet they were largely unknown in training and applications in medical statistics, which instead focused on ascribing informal causal meaning to association measures (such as regression coefficients), subject to equally informal considerations of how well the surrounding evidence warranted that meaning (e.g., Hill, 1965).

Today these formalisms have merged into the vibrant topic area of causal inference and have become part of mainstream methods instruction (e.g., Morgan and Winship, 2007; Rothman et al., 2008). There has in parallel been a subtle conceptual revolution that recognizes causal inference as a prediction problem, the task of predicting outcomes under alternative actions or decisions. The need to consider mutually exclusive actions imposes special structure and special identification issues that parallel problems in destructive testing in engineering, namely the fact that each unit can be observed to fail only under one treatment (e.g., operation at 220 volts rather than 120 volts). Thus we need a basis for inferring from several different groups observed under different treatments to one group that might be subject to any one of the treatments. In product testing this basis is ordinarily provided by random sampling and randomization, but these devices are often unavailable in health and medical studies.

The explication of the predictive structure of causal inference is most explicit in the decision-theoretic formalization of Dawid and colleagues (2004, 2007, 2012). Nonetheless, all formalizations with mainstream acceptance (chiefly potential-outcome models and the parallel decision-theoretic and do-calculus formulations) can be expressed as containing an intervention, action or decision variable (or node) encoding a target "cause" or treatment choice. Any temporally subsequent variable can then be structured as an "effect" or outcome variable, be it mediating or terminating the events of interest.

Within this framework, causal inference can be called "active prediction," forecasting consequences of alternative action or decision sequences (as in Robins, 1986, 1987a, 1987b, and Pearl and Robins, 1995; VanderWeele and Hernán, 2012), albeit sometimes in retrospect, in which case all but one of the alternative actions is counterfactual. The basic idea is traceable at least back to Fisher and Neyman, with their strong emphasis on experimental design and the role of randomization in estimating effects. For nonexperimental inference, it manifests in the importance of *treatment* prediction as opposed or in addition to the outcome prediction of traditional regression analysis (Rubin, 1991), especially in high-dimensional problems (Robins and Wasserman, 2000, sec. 5). That importance stands in contrast to "passive prediction", which forecasts outcomes without formalized intervention or decision nodes, and thus with no formal role for treatment prediction or counterfactuals. It also stands in contrast to most causal inference in practice, where passive prediction methods dominate formal analysis and "treatment" is only a specially labeled regressor that lacks any formal distinction from other regressors.

Potential Outcomes and Missing Data

All statistical procedures rely on assumptions to identify quantities of interest, the most prominent including treatment randomization, random selection, and no or random measurement error (usually conditional on certain observed variables, but these are strong assumptions nonetheless). Of these, treatment randomization stands out as an assumption distinguishing causal inference from passive prediction. This assumption can be reformulated in a number of ways, notably via the isomorphism between the potential-outcome and missing-data formalisms, in which counterfactual outcomes (potential outcomes corresponding to unassigned treatments) are treated as missing values to be imputed (Rubin, 1978, 1991).

Here, causal inference is transformed into a framework of inference under missing data: Given that, for any observation unit, at most one treatment assignment from a list $x_1,\ldots,x_m$ can be made, so at most one component of the potential-outcome vector $\mathbf{Y} = (Y_1,\ldots,Y_m)$ can be observed on the unit. Letting A be the variable coding actual treatment, for a given unit with A=a, x_a is received treatment, and at best only Y_a is observed. All other treatment assignments $x_c \neq x_a$ are counterfactual for the unit, and their corresponding potential outcomes Y_c (or at least some summaries of their distribution) must be imputed for causal inference to proceed.

In ideal physics experiments, the potential outcomes are known functions of treatment determined from established laws or overwhelming evidence, and may be superfluous to list given the function (VanderWeele and Hernán, 2012). For example, if X = temperature applied and Y = liquification of a lead sample, knowing the melting point of lead renders superfluous listing all potential outcomes of the sample at different temperatures. At the other extreme, if no relevant law is known, imputation of missing potential outcomes can be enabled by a missing at random (MAR) assumption. The latter is an immediate consequence of treatment randomization plus random actual-outcome censoring given treatment and baseline covariates. It is an especially delicate imputation problem, however, insofar as no two potential outcomes Y_j and Y_k are ever seen on the same observation unit. Thus randomization/MAR cannot identify any aspect of the joint distribution of $Y_1,\ldots,Y_m$, which has led to debate on whether it is meaningful to even talk of such entities (at least in typical health and social-science applications, where no law is available to determine the joint distribution); e.g., see Dawid and discussants (2000).

The Prognostic View

A useful feature of the missing-data perspective on causation is that it is more oriented toward individual predictions, making it closer to clinical prediction than population-health prediction (epidemiology). It is of course not unique in that regard: The same idea is embodied in the development of passive prediction models for outcomes Y_a, and then re-computation of the

fitted outcomes as prognostic scores by (possibly counterfactually) setting all units to the same treatment value x_j (e.g., Cornfield, 1971). This prognostic scoring constitutes imputation of the expected potential outcome $\mu_j \equiv E(Y_j)$ under treatment x_j.[1]

The idea of prognostic prediction under different treatments has been reborn repeatedly since then, e.g., under the rubric of g-computation, in which the distribution of potential outcomes $p(y_j)$ is computed under different treatment regimes x_j (Robins, 1986, 1987), and again in the do-calculus in which $p(y_j) = p(y|do[x_j])$ (Pearl, 1995; 2009). Arguably, a prognostic structure could be discerned in Neyman's original potential-outcomes model for randomized experiments (Neyman, 1923); the half-century delay in wide dissemination of this approach to formalizing causation may have been because it requires computers and regression software for serious application value.

Ambiguities of Observational Extensions

Formal causal models have produced many valuable insights, and are essential to make sense of mediation analysis and complex interventions (e.g., sequential treatment regimes). Nonetheless, for observational studies (as well as for severely compromised experiments) major challenges to causal inference derive from ambiguous mappings between models and data, and the nonidentification of model parameters from data after ambiguities are resolved.

A preliminary issue that is too often overlooked in observational applications is that potential outcomes (let alone their joint distribution) can be quite ambiguous in the absence of a well-defined *physical* mechanism that determines their assignment index (Greenland, 2005a; Hernán, 2005; VanderWeele and Hernán, 2012). Consider "sex": what does it mean for a female patient to have given treatment "male" versus "female"? The missing variable Y_{male} is undefined without some operational mechanism that transforms the given observation unit from "female" to "male". To avoid this problem, some authors have maintained we should only assume potential outcomes exist when we can specify (perhaps we should add, with a straight face) an assignment mechanism that makes the purported treatment index correspond to values of an action or decision node. Others suggest that extension is possible in theory (again, e.g., when potential outcomes can be determined from known laws) (Pearl, 2009; Shpitser, 2012), although the practical obstacles to extension may be insurmountable in typical health and social-science settings (VanderWeele and Hernán, 2012).

[1]This imputation assumes implicitly a stochastic model for the actual outcomes Y_a, so that potential outcomes are unit-specific expectations $\mu_j = E(Y_a|do[x_j])$ or the corresponding distributions $p(y_j)$ of the Y_j (potential distributions of Y_a) under alternative treatments.

Absent a well-defined treatment-assignment mechanism, it may be sensible and more relevant to switch to another type of observation unit for which such a mechanism can be specified (Greenland, 2005a). For example, to study hiring discrimination, one can change status shown on initial applications from "male" instead of "female" and using a masculine instead of feminine name, to see whether that matters to (say) being invited for an interview. Similar strategies can be applied to study effects of "race", genotypes, and other factors intrinsic to the individual (Kaufman, 2008). Here, the actual treated unit would be the person or group deciding on interviews (not the person applying), and the treatment variable would be their perception of a characteristic, rather than the actual characteristic of the person applying.

Causal Diagrams and Structural Equations

In parallel with potential-outcome and decision-theoretic developments, there gradually emerged a fusion of graphical with causal concepts in the form of path analysis, structural equations, and (eventually) modern causal diagrams (Spirtes et al., 2001; Pearl, 2009). The diagrams themselves are unobjectionable to the extent they merely encode conditional independence assumptions (Dawid, 2010; Greenland, 2010). There are however several formal causal structures that can be identified with these diagrams, with ambiguities and controversies extending from problems with potential outcomes (Robins and Richardson, 2010).

The weakest formulations (which seem to underlie most health and social-science uses) say only that the arrows in the graph represent causation, without much clarity about what that means beyond (a) temporal interpretation of arrows, and (b) the conditional independencies (Markov structure) implied by the graph. The usual condition added for causality is that (c) all confounders are in the graph; nonetheless, it is difficult to understand what a confounder is without a definition of causation attached to the system portrayed. These vagaries may not hinder proper usage, but have compelled theoreticians to promote more precise formulations.

The strongest of these formulations treats structural-equation systems as response schedules giving (what look like, formally) potential outcomes for each endogenous node as a function of its parents. A causal graph then represents a nonparametric structural-equation model (NPSEM), in which a node Y with parents pa(Y) is taken to be determined by a functional relation $Y = f_Y(pa(Y),\varepsilon_Y)$ where ε_Y is exogenous and parent to Y only (Pearl, 2009; Shpitser, 2012). When f_Y is a purely hypothetical function (as opposed to an experimentally tested law), there are serious objections to such formulations, at least if f_Y is considered to supply potential outcomes under different settings for pa(Y) and ε_Y.

One objection is that the NPSEM formulation entails a possibly enormous experimentally nonidentified joint distribution over the all the potential-outcome vectors $\mathbf{Y} = \langle Y_j \rangle$, where j

ranges over values of pa(Y) and Y ranges over graphical nodes. Of at least equal practical concern is the ambiguity of labeling arbitrary variables in pa(Y) as treatments or interventions (discussed above). This problem leads to questions about the realism and utility of such NPSEM formulations (Dawid, 2006; Robins and Richardson, 2010), unless great care is taken to define variables and observation units so that there are acceptably precise meanings for an operator like do[pa(X)] for every variable X in the graph. In response to these objections, one can turn to weaker causal models to match the diagrams, which yield correspondingly less effect identification from a given diagram, or else turn to diagram expansion to match the identification given by the full NPSEM corresponding to the original diagram (Robins and Richardson, 2010).

Compelling vs. Plausible Assumptions, Models, and Inferences

If one has established acceptably precise and relevant treatment units and variables, precise causal models can be formulated. In practice, causal models typically encode numerous assumptions (constraints) beyond those implicit in their definitions of cause and effect. Because those further assumptions turn out to be essential for identification of effects, it is important to distinguish compelling assumptions from those that are merely plausible. This is an obvious point worth belaboring because it is so often forgotten in theory *and* practice.

Evidentially compelling assumptions are supported by a body of evidence taken to refute all of what might have otherwise been plausible alternatives. They are thus felt safe to treat as givens. Consequently, a model may be labeled as compelling if it entails only compelling assumptions, or is flexible enough to approximate such a model. Similarly, an argument is compelling if all its assumptions are compelling and its deductions are sound.

Study-design features are often taken as compelling evidence for strict constraints (dimension-reducing assumptions). For example, baseline (pre-treatment) matching on an exogenous covariate (e.g., age) when constructing a study cohort can compel the strict (sharp-null) assumption that the covariate is unassociated with treatment and thus unconditionally not a confounder of any treatment effect on the cohort. On the other hand, background mechanistic information may only compel relatively vague constraints, although those may involve strict inequalities. As an example, monotone dose-response (e.g., monotonicity of the plot of μ_j against x_j) is often taken to be compelling, such as the relation of carbon monoxide concentration to death by asphyxiation.

An assumption is sometimes said to be warranted or credible or justified if there is some justification for use, such as supporting evidence that refutes explicit competitors. A plausible assumption may have no specifically supporting evidence or justification, and may simply be

one among many unrefuted competitors. Plausibility often stems only from entailment of compelling consequences, or from some heuristic or esthetic advantages such as simplicity. In that case it is important to remember that many other assumptions will have the same source of plausibility. For example, linear dose-response (i.e., linearity of the X effect on Y, linearity of the plot of Y_j against x_j) is often plausible and so is common in everyday "working models". Linearity implies strict monotonicity, so when the latter is found compelling, the plausibility of linearity is enhanced; but the same can be said of many simple alternatives such as loglinearity.

To use plausible assumptions safely, the consequences of their violation (under plausible alternatives) need to be understood in some detail to see if they are acceptable within our estimation task. Linearity may be relatively harmless for population inferences when (a) strict monotonicity holds, and (b) one restricts inferences to what happens on average to Y as X varies (i.e., the marginal effect, which is the slope across the plots of Y_j against x_j averaged over the distribution of the potential-outcome vector **Y** in the source population of units). But if we assume linearity when responses are in fact nonlinear, this average slope can lead us into costly mistakes in predicting treatment effects (e.g., at high doses). This is so even if our average-slope estimate is from a perfect randomized trial.[2] Furthermore, even if individual responses are linear, heterogeneity of the individual slopes may also lead to costly mistakes in individual (clinical) predictions of treatment effects, especially if the slopes change sign across individuals (e.g., if the average slope is negative indicating a benefit on average, but the treatment is harmful in some individuals).

A model may be labeled as plausible if it entails *only* plausible assumptions, or is flexible enough to approximate such a model. It is a logical fallacy to claim plausibility of a model from its entailment of plausible or compelling consequences; nonetheless, if one quantifies plausibility with subjective probability, it is logical to increase the plausibility of a model upon observing one of its logical consequences or upon refutation of alternatives (e.g., see Adams, 1998). Unfortunately, purely logical relations (even probabilistic ones) do not seem to get us very far in statistical applications, which may explain (although not excuse) why logic is so neglected in statistical training.

As an example, suppose 20 assumptions $A_1,\ldots,A_{20}$ entail a conclusion B. This means that we should assign B no less than the probability of $A_1\&\ldots\&A_{20}$, with B certain (probability 1) if we are certain of every A_j. It is then tempting to think that B must be probable if evidence leads us to assign each A_j a probability of 0.95, a common interpretation of "very probable". But if these

[2]For example, suppose in a cohort the death risks (expected potential outcomes) at doses X=0, 1, 2 were 0.4, 0.1, and 0.1, and we randomize the cohort among these three doses. Then the expected fitted values from a linear model would make X=2 appear the most beneficial dose even though full benefit is achieved at X =1. If instead the death risks at X=0, 1, 2 were 0.4, 0.4, and 0.1, the expected fitted values from a linear model would make X=1 appear beneficial even though it is not.

are subjectively independent assumptions, this information alone does not imply B should have probability greater than $0.95^{20} = 0.36$, with an even lower minimum if (as is usual) B is not perfectly entailed by $A_1,\ldots,A_{20}$. Thus, even if we assign "very high" probability to each of twenty independent supporting pieces of evidence A_j (which sounds very impressive) and B can be deduced logically from the twenty, it does not follow that B must be highly probable. For the latter conclusion, the *conjunction* of the assumptions must have high probability, which might not be the case if each assumption is subject to independent uncertainties or unknown errors.

Given that pure logic and probability may not take us far when given only independent bits of uncertain assumptions or evidence, causal inference often turns to highly interdependent and coordinated sets of assumptions in the form of mechanisms (mechanistic models). Although these models can be treated conveniently as singular hypotheses or assumptions, such treatment should not obscure their compound nature, especially when evaluating their credibility. In particular, a mechanistic model, by virtue of its elegance or other non-evidential properties, may induce a conjunction fallacy in which the model attains higher probability than any of its assumptions (Tversky and Kahneman, 1983).

Even if we agree that a particular mechanistic model is correct or compelling, the model's scope and impact may be far more narrow than believed, at least when used to argue against causation (which itself is a very large compound hypothesis embracing all causal mechanisms). Take the sharp null hypothesis that cell phones cannot cause brain cancer in humans. Clearly, this hypothesis has never been tested by a "black box" experiment in which people are randomized to different exposure levels. Nonetheless, Shermer (2010) argued that this null hypothesis is compelling (in fact, certain), based on the mechanical argument that cell-phone microwaves (CPMs) are nonionizing, and nonionizing radiation cannot alter chemical bonds including those in cellular DNA. The underlying physics is a mechanistic model, which is being offered as a compelling assumption that leads to a compelling conclusion that a particular etiologic pathway is not operating. This physics is uncontroversial, and Shermer uses that fact as part of an appeal for the general causal null hypothesis.

Shermer's fallacy comes in applying the physics to the biology. To do so, he inserts the assumption that an exposure (such as CPM) can elevate cancer risk only by direct breakage of DNA. This assumption is not even plausible to some cancer biologists because (among other things) most carcinogenesis does not involve ionizing radiation, and may not even require direct DNA damage; merely impeding DNA repair mechanisms or gene expression may suffice. Several experimental studies have reported that CPMs can in fact influence DNA (e.g., see list in Slesin, 2008). Furthermore, the DNA argument applies only to tumor initiation; it does not apply to tumor promotion, which can be affected by many factors including metabolic processes that favor neoplastic over normal cells. In this regard it may be noted that CPMs have been

experimentally reported to affect brain metabolism (Salford et al., 2003; Volkow et al., 2011). Consequently, the null hypothesis is not compelling to everyone – at best, the physics only argues against a certain narrow class of mechanisms (involving direct DNA damage) that are alternatives to the null. By ruling out this class, the physics thus should increase the plausibility of the null for those who accept the mechanistic argument, but cannot render the null certain until all other mechanisms are ruled out as well. Thus, there are far too many mechanisms left open by current evidence to even claim the physics should render the null a warranted assumption. Given that randomized trials are infeasible, we are left with only epidemiologic evidence converging on no association (e.g., Frei et al., 2011) to support the notion that no such mechanism is making an important contribution to population risk.

Formal methods for incorporation of narrow or uncertain mechanistic information into causal inference as yet seem limited compared to methods for incorporating data (whose evidential value can be measured and added to the inferential process via loglikelihoods or penalties). The recent surge of interest in mechanism representations such as sufficient-cause models (VanderWeele, 2012) may however open an avenue for such incorporation (e.g., perhaps via introduction of prior distributions over sufficient causes). Conversely, these representations allow testing of certain mechanistic hypotheses (e.g., about "interactions") with epidemiologic data (Rothman et al., Ch. 16; Berzuini and Dawid, 2012; VanderWeele, 2012) – but only under identifying assumptions that biases are negligible. Under compelling assumptions, epidemiologic data *alone* rarely identify the global causal-null hypothesis, let alone interactions or specific mechanisms (Greenland, 2005b, 2009, 2010).

Nonidentification and the Curse of Dimensionality

It seems natural to ask why we don't always use compelling models instead of merely plausible models. One reason is the controversy surrounding most certainty claims (as just exemplified for cell phones); another common reason is lack of data to justify such claims. The result is that a high degree of flexibility is needed to produce constraints that might be accepted by all involved as "compelling" in a situation.

Consider approximation of multivariate monotonicity without use of additivity or linearity, as in nonparametric regression. Such flexibility will demand at least several terms for each variable alone and an explosion of products among them, along with complex constraints on the terms. With enough variables, conventional modeling rapidly hits the "curse of dimensionality" in which small-sample artifacts begin to plague conventional fits, even with what appear to be large data sets (large numbers overall, but not enough to support maximum likelihood for the model). To some extent these artifacts can be minimized and performance greatly improved by

use of priors or penalty functions (shrinkage) on model parameters; e.g., see Greenland, 2008, p. 525-527 for citations to some examples and studies. Thus penalized estimation is arguably a better norm for practice (especially since conventional methods are a special case of penalized methods with zero penalty). Nonetheless, penalized estimation used alone retains the same "effective" dimensional limits; it simply allows more flexible allocation of degrees of freedom.

The algorithmic-modeling (machine-learning) literature tackles the dimensionality problem directly via intensive simulation (Breiman, 2001; Hastie et al., 2009) and enables consistent estimation of many properties of high-dimensional distributions, as long as those properties are identified by given constraints. It has been applied to good effect in causal-inference problems both for weight estimation (McCaffrey et al., 2004; Lee et al., 2011) and more directly (van der Laan and Rose, 2011). But in common observational settings, compelling assumptions may no longer identify targeted causal quantities, at least if (as in classical frequentist inference) we are limited to strict constraints. Put another way, relaxation of assumptions to move from the plausible to the compelling can lead to interval estimates whose width increases without bound.

On the other side, re-introducing merely plausible assumptions (e.g., linearity, additivity) to impose identification can create inferences that are seriously biased and overconfident, especially in failing to incorporate the uncertainty surrounding those assumptions (Greenland, 2005b; Richardson et al., 2010). This caution applies whether the assumptions are within a fully parametric, semiparametric, or nonparametric framework. For example, most causal-modeling methods assume randomization (often couched as nonconfounding or ignorability) conditional on certain observations, which nonparametrically identifies effects via conditional permutation tests. But that assumption has no compelling support in typical epidemiologic contexts, and that lack often leads to considerable controversy about the meaning, implication, and value of the test results.

It should be appreciated too that although an identifying assumption may look simple, it may entail an enormous number of important conditions and thus be far stronger than its form makes it seem. For example, conditional randomization/no-confounding/ignorability can written simply as independence of treatment X and potential outcomes Y_j given confounders. But when there are many unmeasured or poorly measured potential confounders, or there are many measured potential confounders, the assumption that any fitted propensity score or other empirical conditionalization produces the needed independency comes down to assuming that the many omitted terms are nonconfounding (unrelated to both X and the Y_j at once).

To deal with recognized nonidentification, most statistical treatments turn to sensitivity analyses in which identifying assumptions are varied. Hopefully these variations are over

plausible assumptions, since without identification a complete sensitivity analysis will only display the range allowed for the target parameter by the fixed part of the model and any bounds on parameter exploration (Vansteelandt et al., 2006). This type of analysis can be used to see whether plausible violations could be responsible for observed associations or masking important effects. From considerations of plausibility it is but one more step to impose prior distributions over the identifying assumptions (suitably parameterized) and proceed with Bayesian or similar analyses (e.g., Leamer, 1974; Eddy et al., 1992; Phillips, 2003; Lash et al., 2009; Greenland, 2005b, 2009; Gustafson, 2005; Turner et al., 2009). Nonetheless, these variations or priors (or variations on priors) will again face a curse of dimensionality in large problems: Unless strict constraints are imposed, there will be too many sensitivity dimensions to explore let alone set priors on effectively (e.g., see the general confounding sensitivity formulas in VanderWeele and Arah, 2011).

Identification in Practice

As a result of formal nonidentification, most analyses remain a pastiche of highly formalized and completely informal activities that induce an informal sort of identification. These activities often follow a pattern along these lines:

a) assume a relatively simple, plausible identifying model (usually from a *very* short list that will go unquestioned in the context: linear, log-linear, etc.);
b) go through the mechanized and often highly elaborate fitting process;
c) add conditionals that amount to "if our model is correct…" to the resulting formal statistical inferences;
d) pull back from those inferences with cautions that amount to "our model may not be correct", along with speculations (often highly biased) about the likelihood and consequences of violations of single assumptions (which may not reflect accurately the likelihood and consequences of multiple violations);
e) conclude by reasserting the model-based inferences in step (c) as if they were observations, not inferences ("we observed an/no effect of X on Y").

Of course (e) is not necessary and savvy researchers water it down to the extent referees and editors will allow. But in step (d), particular assumptions may be defended or criticized based on appeals to other studies or mechanisms.

Mechanistic hypotheses are difficult to calibrate or weight in the inference, and as a result are often used overconfidently. At the extremes, authors may argue as if mere description of a plausible mechanism demonstrates causation, or as if the failure to describe a plausible mechanism supports or even proves no causation. Thus the assumption that cell phones cannot

produce direct DNA damage is based on "physics," which is code for a plethora of studies establishing laws (theories accepted as facts) taken to prove that CPMs do not have sufficient energy for direct chemical effects. Again, such invocation does not exclude other causal pathways, and so does not justify adopting strictly the null hypothesis of no CPM effect, although it may shrink our prior distribution toward that null.

Analogous overconfidence in empirical evidence arises when studies are cited as if their results are more certain or demonstrative than they turn out to be when scrutinized closely. The latter problem is indicated by statements that studies "showed" or "found" some result (whether an association or no association). Inevitably, examination of the cited sources (whether experimental or epidemiological) reveal nothing to warrant such certainty upon considering interval estimates and study limitations. Yet the overconfidence placed in other reports often contributes to overconfident conclusions in the current report, which in turn are cited overconfidently in further reports. And all the reports were overconfident from the use of arbitrary identifying constraints.

In a parallel fashion, there is often overconfident dismissal of still other reports, usually based on limitations inherent in the study design and execution. Typically, a limitation introduces the possibility of a bias mechanism (e.g., lack of randomization opens the door to confounding mechanisms), but rarely does it imply that the mechanism operated, and still more rarely would we know the extent of bias it produced. For example, collection of data from interviews after the outcome opens the door to differential recall (dependence of recall on outcome, "recall bias"), but that fact alone does not imply that this phenomenon had any important impact, nor does it imply (as commonly assumed) that it pushed estimates away from the null (Drews and Greenland, 1990). A design limitation leaves open bias pathways or uncertainty sources, just as the corresponding design strength removes those pathways. Thus post-outcome recall leaves open bias pathways operating through outcome effects on response, while baseline data collection excludes that pathway; but the general nonidentitication problem remains.

One way of breaking the cycle of citation overconfidence (whether excessive credence or unjustified dismissal) is to enter all collected data into a single analysis, treating design features as regressors (meta-regression). But it is extremely difficult to obtain those data, and when they are obtained, conventional analyses can amplify another source of overconfidence: The computed statistical precision of the final estimate. With large enough numbers and forced identification, the standard errors will shrink to the point that they represent a minor component of warranted uncertainty.

Sensitivity analysis can provide an antidote to this statistical overconfidence by showing how apparently precise inferences may evaporate under certain model expansions. But it can be misused in the opposite extreme to claim that no inference is possible, even when diverse

evidence points to a conclusion. One approach to the middle ground employs a greatly expanded model that incorporates all the known major sources of nonidentification, and displays results from a system of plausible priors (Greenland, 2005b, 2009; Gustafson, 2005; Lash et al., 2009). But, like other approaches, these expansive analyses retain the difficulties of incorporating mechanistic information, and can always be criticized for their inevitable failure to explore all imaginable dimensions or plausible priors for the problem.

Identification and Bounded Rationality

For better or worse, we make inferences and decisions when there are too many uncertain dimensions to be evaluated by data or deduction. It seems that some persons are much better at this than others, as evidenced by how well they do subsequent to their decisions (given their background). We thus might look to behavioral studies to discover heuristics for dealing with the dimensionality problem.

Those studies are justly celebrated for revealing the severe and often damaging biases that enter into common heuristics (Gilovich et al., 2002; Kahnemann et al., 1982). Notable among them is overconfidence, e.g., from overvaluation of evidence or opinions (one's own or those of trusted parties), which leads to adoption of unfounded strict constraints. Yet, as described in theories of bounded utility and rationality (Simon, 1991; Gigerenzer and Selten, 2002), and recognized in common cautions against "overthinking", rational decisions require constraints on our deductive activity; i.e., we must rationally allocate or constrain analysis (called satisficing by Simon, 1956, and type-II rationality by Good, 1983). Indeed, the most common rational objections to new methods that I have heard from epidemiologists are of the form "Is this technique really worth the effort?" And often (but by all means not always) the only honest and well-informed answer is that it depends on the cost of applying the method, and the risk of error in applying the method versus the risk of error from failing to apply the method.

Adoption of an initial model represents a tentative commitment to constrain analysis within the confines of the model. Where do such initial constraints come from? Values (including values assigned to traditions, as well as effort costs) are one source. It has long been recognized that values enter into methodologies, including those for scientific inference (Kuhn, 1977; Poole, 2001), and they are not hard to discern in modeling. Consider again linearity: Its adoption represents a commitment to not consider nonlinear models unless persuasive evidence of nonlinearity comes to our attention. Absent compelling empirical support (which would obviate the study under analysis), linear models are still rationalized based on inter-related values of simplicity, parsimony, and transparency, plus conformity to tradition (in order to minimize criticism and questioning of methodology).

The role of values (as opposed to direct rationalization) becomes more obvious and dubious in medical-evidence ratings, which simply declare some studies stronger evidence than others based on gross classification schemes, without attention to finer details that could undermine the rational basis for the declaration in specific cases. A simple example is randomized trials vs. nonrandomized cohort studies. Medical journals automatically classify trials as a better "level of evidence" than observational cohorts, thus leading themselves and readers to value trials more than cohorts. Yet it is quite possible for a trial to be more misleading than an observational cohort. This can occur in many ways, even with neither study being biased in the narrow sense of statistical bias or internal invalidity.

As an example, typical drug-approval trials are too small and too short to reveal rare or long-term side effects, and may be done in selected low-risk populations. As a result, they are prone to give an impression of safety that may be dispelled in a study large health-care data bases. As another example, suppose the trial took place among low-risk volunteers but the cohort comprised high-risk patients and high-risk patients will receive the treatment in practice. The effect could be quite different in low-risk and high-risk populations; such a difference would constitute a bias present in the trial and absent in the cohort, and might even reverse their true evidential value. The lack of provision for such reversal in current medical evidence schemes illustrates the hazards of expedient simplifications for evidence evaluation, even when simplifications are needed on type-II rational grounds.

Conclusion

Many of the problems discussed here apply to scientific inference in general, whether labeled causal or not, especially problems of identification and evidence synthesis. There is a huge literature on methodologies for combination of evidence, so large that I cannot even begin to provide representative citations, let alone claim any expertise. Issues of combining highly disparate evidence arises in legal and management as well as scientific settings, and may provide valuable clues for improving causal-inference methodology (at least for those not averse to connecting inference and decision-making). But as yet none of those methodologies has become prominent in health-science research or the causal-inference literature, despite some overlap between evidence-synthesis and causal-inference research (e.g., Hepler et al., 2009; Turner et al., 2009).

The current lack of a received methodology for actual causal inference (with synthesis of diverse types of evidence) should be unsurprising, given the difficulty of formalizing such a complex (but essential) task, and user limitations in applying complex formalisms. In light of this situation, it has been suggested that causal inferences might often just as well be left informal,

at least whenever causal effects are far from identified in any realistic model (Greenland, 2010). After all, despite decades of formal developments, there is much wisdom to be had in informal guidelines for inference such as the Hill (1965) considerations and later refinements (e.g., Susser, 1991), as long as they are not taken as codified dogmas or criteria (Phillips and Goodman, 2004).

None of this is to dismiss the utility of formal causal models. We need these models to give precise causal meaning to the associations we offer as estimated effects. That need becomes particularly acute with complex treatments, longitudinal interventions, and mediation analysis. But the primary challenge in producing an accurate causal inference or effect estimate is most often that of integrating a diverse collection of ragged evidence into predictions to an as-yet unobserved target. This process does not fit into formal causal-inference methodologies currently in use, which by and large assume we have homogeneous observations from the target (embodied in phrases like "imagine N independent identically distributed copies of X"). Thus, while theory and methods for causal modeling have come a long way in past 40 years, they still have a very long way to go before they approach a complete system for causal inference.

Acknowledgment: I wish to thank Charles Poole, Katherine Hoggatt, Jay Kaufman, and Tyler VanderWeele for helpful comments on this chapter.

References

Adams, E. W. (1998). *A Primer of Probability Logic*. Chicago: CSLI Publications/Univ. of Chicago Press.

Berzuini, C. and Dawid, A.P. (2012). Mechanistic inference from epidemiological data. In this volume.

Breiman, L. (2001). Statistical modeling: The two cultures. Statistical Science, 16, 199-215.

Cornfield, J. (1971). The University Group Diabetes Program: A further statistical analysis of the mortality findings. Journal of the American Medical Association, 217, 1676-1687.

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

Dawid, A. P. (2004). Probability, causality and the empirical world: A Bayes-de Finetti-Popper-Borel synthesis. Statistical Science 19, 44-57.

Dawid, A. P. (2007). Counterfactuals, hypotheticals and potential responses: A philosophical examination of statistical causality. In Causality and Probability in the Sciences, Texts in Philosophy, Vol. 5, ed. F. Russo and J. Williamson, pp. 503–32. College Publications, London. Available at http://www.ucl.ac.uk/statistics/research/pdfs/rr269.pdf

Dawid, A. P. (2010). Beware of the DAG! In: Proceedings of the NIPS 2008 Workshop on Causality. Guyon, I., Janzing, D. and B. Scholkopf, B., eds. Journal of Machine Learning Research, Workshop and Conference Proceedings, 6, 59-86. Available at http://jmlr.csail.mit.edu/proceedings/papers/v6/dawid10a/dawid10a.pdf

Dawid, A. P. (2012). The decision-theoretic approach to causal inference. In this volume.

Drews, C. and Greenland, S. (1990). The impact of differential recall on the results of case-control studies. International Journal of Epidemiology, 19, 1107-1112.

Eddy, D.M., Hasselblad, V. and Shachter, R. 1992). Meta-analysis by the confidence profile method. New York: Academic Press.

Frei, P., Poulsen, A.H., Johanssen, C., Olsen, J.H., Steding-Jessen, M.and Schuz, J. (2011). Use of mobile phones and risk of brain tumours: update of Danish cohort study. *British Medical Journal*, 343, in press.

Gigerenzer, G. and Selten, R. (2002). *Bounded Rationality: An Adaptive Toolbox*. Cambridge: MIT Press.

Gilovich T, Griffin D, Kahneman D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press.

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

Greenland, S. (2005a). Epidemiologic measures and policy formulation: Lessons from potential outcomes (with discussion). Emerging Themes in Epidemiology (online journal) 2:5. Available at http://www.ete-online.com/content/pdf/1742-7622-2-5.pdf

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. (2008). Variable selection and shrinkage in the control of multiple confounders (invited commentary). American Journal of Epidemiology, 167, 523-529; erratum: 1142.

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

Greenland, S. (2010). Overthrowing the tyranny of null hypotheses hidden in causal diagrams. Ch. 22 in: Dechter, R., Geffner, H., and Halpern, J.Y. (eds.). *Heuristics, Probabilities, and Causality: A Tribute to Judea Pearl*. London: College Press, 365-382. Available at http://intersci.ss.uci.edu/wiki/pdf/Pearl/22_Greenland.pdf

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.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York.

Hepler, A. B., Dawid, A. P. and Leucari, V. (2007). Object-oriented graphical representations of complex patterns of evidence. Law, Probability & Risk 6, 275-293.

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

Hill, A.B. (1965). The environment and disease: association or causation? Proceedings of the Royal Society of Medicine, 58, 295–300.

Kahneman, D., Slovic, P. and Tversky, A. (1982). *Judgment under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press.

Kaufman, J,S. (2008). Epidemiologic analysis of racial/ethnic disparities: Some fundamental issues and a cautionary example (with discussion). *Social Science and Medicine*, 66,1659-1680.

Kuhn, T.S. (1977). Objectivity, value judgment, and theory choice. In: *The Essential Tension*, ed. TS Kuhn. Chicago: Univ. Chicago Press, 320–43.

Lash TL, Fox MP, Fink AK. *Applying Quantitative Bias Analysis to Epidemiologic Data*. Boston, MA: Springer Publishing Company; 2009.

Leamer, E.E. (1974). False models and post-data model construction. Journal of the American Statistical Association, 69, 122–131.

Lee, B.K., Lessler, J. and Stuart E.A. (2011) Weight trimming and propensity score weighting. PLoS One 6(3): e18174. doi:10.1371/journal.pone.0018174. Available at http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0018174

McCaffrey, D.F., Ridgeway, G., Morral, A.R. (2004) Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological Methods, 9, 403–425.

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

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. English translation in: Statistical Science 1990, 5, 465–480.

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

Pearl, J. (2009). Causality: Models, Reasoning, and Inference. 2nd ed. Cambridge University Press, New York.

Pearl, J. and Robins, J.M. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In: P. Besnard and S. Hanks, eds. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. San Mateo, CA: Morgan Kaufmann Publishers, 444-453.

Phillips, C.V. (2003). Quantifying and reporting uncertainty from systematic errors. Epidemiology, 14, 459-466.

Phillips CV, Goodman KJ (2004). The missed lessons of Sir Austin Bradford Hill. Epidemiol Perspect Innov 1:3. (online journal) doi:10.1186/1742-5573-1-3

Poole, C. (2001). Causal values. Epidemiology, 12, 139-141.

Robins, J.M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period: Applications to control of the healthy worker survivor effect. Mathematical Modeling 7, 1393-1512.

Robins, J.M. (1987a). Addendum to: A new approach to causal inference in mortality studies with a sustained exposure period. Computers and Mathematics with Applications, 1987, 14, 917-945; errata 1987, 18, 477.

Robins, J.M. (1987b). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. Journal of Chronic Disease (40, Supplement), 2:139s-161s.

Robins, J.M. and Wasserman L. (2000). Conditioning, likelihood, and coherence: A review of some foundational concepts. Journal of the American Statistical Association, 95, 1340-1346.

Robins, J.M. and Richardson, T.S. (2010). Alternative graphical causal models and the identification of direct effects. In: Shrout, P., ed. Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures. New York: Oxford University Press.

Rothman, K.J., Greenland, S. and Lash, T.L. (2008). Modern Epidemiology, 3rd ed. Philadelphia: Lippincott-Wolters-Kluwer.

Rubin, D.B. (1978). Bayesian inference for causal effects: the role of randomization. Annals of Statistics, 6, 34–58.

Rubin, D.B. (1991). Practical implications of modes of statistical inference for causal effects, and the critical role of the assignment mechanism. Biometrics, 47, 1213–1234.

Salford LG, Brun AE, Eberhardt JL, Malmgren L, Persson BRR (2003). Nerve cell damage in mammalian brain after exposure to microwaves from GSM mobile phones. Environmental Health Perspectives, 111, 881-883.

Shermer, M. (2010). Can you hear me now? Scientific American, October. Available at http://www.michaelshermer.com/2010/10/can-you-hear-me-now/

Shpitser, I. (2012). Introducing determinism: the language of structural equations models. In this volume.

Simon, H. A. (1956). Rational choice and the structure of the environment. Psychological Review, Vol. 63 No. 2, 129-138.

Simon, Herbert (1991). Bounded rationality and organizational learning. Organization Science 2, 125–134.

Slesin, L. (2008). Science magazine gets it wrong on DNA breaks. Microwave News, 28, 1-3. Available at http://www.microwavenews.com/docs/mwn.11%2810%29-08.pdf

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

Susser, M. (1991). What is a cause and how do we know one? A grammar for pragmatic epidemiology. American Journal of Epidemiology, 133, 635-648.

Turner, R.M., Spiegelhalter, D.J., Smith, G.C.S. and Thompson, S.G. (2009). Bias modeling in evidence synthesis. Journal of the Royal Statistical Society, Series A, 172, 21-47.

Tversky, A. and Kahneman, D. (1983). Extension versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review, 90, 293–315.

van der Laan, M.J. and Rose, S. (2011). Targeted Learning: Causal Inference for Observational and Experimental Data. New York, Springer.

VanderWeele, T.J. (2012). The sufficient cause framework in statistics, philosophy and the biomedical and social sciences. In this volume.

VanderWeele, T.J. and Arah, O.A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments and confounders. Epidemiology, 22, 42-52.

VanderWeele, T.J. and Hernán, M.A. (2012). Causal effects and natural laws: towards a conceptualization of causal counterfactuals for non-manipulable exposures. In this volume.

Vansteelandt, S., Goetghebeur, E., Kenward, M.G. and Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. Statistica Sinica, 16, 953-980.

Volkow ND, Tomasi D, Want G, et al. (2011). Effects of cell phone radiofrequency signal exposure on the brain glucose metabolism. JAMA, 305, 808-814, 828-829.