Taylor & Francis
Taylor & Francis Group

Critical Reviews in Toxicology

ISSN: 1040-8444 (Print) 1547-6898 (Online) Journal homepage: http://www.tandfonline.com/loi/itxc20

Do causal concentration–response functions exist? A critical review of associational and causal relations between fine particulate matter and mortality

Louis Anthony (Tony) Cox Jr

To cite this article: Louis Anthony (Tony) Cox Jr (2017): Do causal concentration–response functions exist? A critical review of associational and causal relations between fine particulate matter and mortality, Critical Reviews in Toxicology, DOI: 10.1080/10408444.2017.1311838

To link to this article: http://dx.doi.org/10.1080/10408444.2017.1311838

CRITICAL REVIEWS IN TOXICOLOGY, 2017
http://dx.doi.org/10.1080/10408444.2017.1311838

Taylor & Francis
Taylor & Francis Group

REVIEW ARTICLE

Do causal concentration–response functions exist? A critical review of associational and causal relations between fine particulate matter and mortality

Louis Anthony (Tony) Cox Jr

Cox Associates, Denver, CO, USA

ABSTRACT

Concentration–response (C–R) functions relating concentrations of pollutants in ambient air to mortality risks or other adverse health effects provide the basis for many public health risk assessments, benefits estimates for clean air regulations, and recommendations for revisions to existing air quality standards. The assumption that C–R functions relating levels of exposure and levels of response estimated from historical data usefully predict how future changes in concentrations would change risks has seldom been carefully tested. This paper critically reviews literature on C–R functions for fine particulate matter (PM2.5) and mortality risks. We find that most of them describe historical associations rather than valid causal models for predicting effects of interventions that change concentrations. The few papers that explicitly attempt to model causality rely on unverified modeling assumptions, casting doubt on their predictions about effects of interventions. A large literature on modern causal inference algorithms for observational data has been little used in C–R modeling. Applying these methods to publicly available data from Boston and the South Coast Air Quality Management District around Los Angeles shows that C–R functions estimated for one do not hold for the other. Changes in month-specific PM2.5 concentrations from one year to the next do not help to predict corresponding changes in average elderly mortality rates in either location. Thus, the assumption that estimated C–R relations predict effects of pollution-reducing interventions may not be true. Better causal modeling methods are needed to better predict how reducing air pollution would affect public health.

ARTICLE HISTORY
Received 12 October 2016
Revised 23 March 2017
Accepted 23 March 2017

KEYWORDS
Causality; concentration–response functions; C–R; PM2.5; manipulative causality; predictive causality; Bayesian networks; *randomForest*

Table of contents

CONTACT Louis Anthony (Tony) Cox tcoxdenver@aol.com Cox Associates, 503 Franklin Street, Denver, CO 80218, USA

Introduction: causal vs. associational concentration–response relations

A concentration–response (C–R) curve shows levels of adverse health responses in exposed populations on its vertical axis and levels of ambient concentrations of a pollutant on its horizontal axis. Such curves, which are usually upward-sloping, have been widely used to predict the public health impacts of proposed reductions in air pollutants (Schwartz et al. 2002; Pope et al. 2015). These predictions are made by assuming that reducing air pollution will reduce adverse health effects, moving both exposure and response variables leftward and downward along the C–R curve. Thus, a C–R curve is commonly given both of the following two interpretations:

Associational interpretation. A C–R curve shows the estimated response (R) for each estimated level of exposure concentration (C), based on historical data. The association between C and R is typically described by a regression model. Such a model may also include trends, season, weather, and other variables, including co-pollutants (Schwartz et al. 2002). We call this the *associational* interpretation of a C–R curve. It describes C–R associations between levels of C and levels of R in historical data. Common measures of C–R associations include relative risk (RR), the regression coefficient for C as a predictor of R, and quantities derived from them such as the odds ratio (OR), population attributable risk, etiologic fraction, or global burden of disease estimates.

Causal interpretation. The C–R curve shows what the estimated response *would become* if the exposure concentration were to be fixed at different levels. Thus, it also shows how the response would change if concentration were changed.

As illustrated in detail in Table 1 later, these two interpretations are usually conflated, leading to causal interpretations of associations. For example, causal claims such as that "The magnitude of the *association* suggests that controlling fine particle pollution *would result* in thousands of fewer early deaths per year" (Schwartz et al. 2002, emphases added) are very common in the epidemiological literature on air pollution health effects, and in regulatory risk assessments based on this literature, even though, as discussed next, there is no necessary relation between the magnitude or direction of a historical C–R association and the effects on R of reducing C.

In principle and in practice, there may be no single curve satisfying both interpretations. For example, suppose that a positive association were to be found between daily consumption of aspirin and heart attack risk in an elderly population, perhaps because elderly people at higher risk of heart attacks are more likely to be prescribed an aspirin regimen to reduce that risk. Clearly, such an empirical association between exposure (or consumption) levels and risk levels would imply nothing about how *changing* daily consumption of aspirin would *change* future heart attack risks. The data only address associations between historical levels, which do not reveal the impacts of future changes. The historical association between their levels may be positive and the future relationship between changes in their levels negative, so that reducing aspirin consumption would increase risk, even though historical levels of aspirin consumption and risk are positively correlated. More generally, finding a positive C–R association in historical data does not necessarily imply anything about how changing C would change R. It is *not* generally true that if a C–R model describes past data values for C and R, then it also predicts how changing exposure concentration from a current level to a new level will change the average response.

Example: C–R associations do not necessarily provide valid causal predictions

To illustrate, consider a simplified setting in which daily mortality rate, R and average daily exposure concentration, C have been measured for several years and the following associational C–R model perfectly fits the data:

$$R = \mathrm{C} + 50 \qquad \text{(Model 1: Associational C–R relation)}$$

That is, each additional unit of exposure concentration is associated with an additional unit of daily mortality, which on these scales corresponds to a 2% increase above the zero-exposure baseline mortality rate of 50. Do these data imply or suggest that reducing C would reduce R? Not necessarily, because an associational model does not necessarily represent causal mechanisms (McClellan 2016; Moolgavkar 2016). Other variables may affect how or whether R changes when C changes. For example, suppose that an unmeasured third variable, T, such as minimum daily temperature, affects C and R as described by the following structural equations:

$$C = 50\text{–}0.5T, \text{ for } 0 \le T \le 100$$
$$R = 150\text{–}C\text{–}T, \text{ for } 0 \le C + T \le 150$$
$$\text{(Model 2: Causal C–R relation)}$$

These equations have the following explicit causal interpretation: if the value of a variable on the right side of an equation is changed, then the value of the dependent variable on the left side will change to restore equality. Thus, changes propagate from right to left through these equations. The causal C–R relation is then very simple: decreasing C by one unit via an exogenous intervention would *increase* R by one unit (since the two are *negatively* related, with a C–R coefficient of -1 for the causal impact of changes in C on changes in R). The associational C–R relation is simply Model 1. (Solving the first equation for T (yielding $T = 100 - 2C$) and substituting into the second equation to eliminate T yields the reduced-form C–R equation: $R = C + 50$.) Model 1 correctly reveals that days with one unit less of C have historically been associated with one unit *less* of R (i.e. the two are *positively* associated), even though Model 2 correctly reveals that reducing C by one unit would *increase* R by one unit. Thus, the causal and associational C–R relations are quite different. The associational model would be entirely valid for predicting how many deaths would occur on days with different exposure concentrations in the absence of interventions, but only the causal model can be used to correctly predict how changing C would change R. In reality, of course, many C–R models include temperature, but other unmeasured, unmodeled, or mis-modeled variables can drive wedges between associational and causal C–R relations, so care should be taken to distinguish between these potentially

Table 1. Examples of association and causation being conflated in the literature on PM2.5 health effects.

Interpretation in literature	Comments
"We observed statistically significant and robust *associations* between air pollution and mortality … these results suggest that fine-particulate air pollution, or a more complex pollution mixture associated with fine particulate matter, *contributes to* excess mortality in certain U.S. cities." Dockery et al. (1993)	Associations do not suggest a contribution to excess mortality unless they are causal.
"The magnitude of the *association* suggests that controlling fine particle pollution *would result* in thousands of fewer early deaths per year." Schwartz et al. (2002)	Associations do not allow prediction of results from changes in exposure concentrations unless the associations represent manipulative causal relations.
"We examined the *association* between PM(2.5) and both all-cause and specific-cause mortality … Our findings describe the magnitude of the *effect* on all-cause and specific-cause mortality, the modifiers of this association, and suggest that PM(2.5) may pose a *public health risk* even at or below current ambient levels." Franklin et al. (2006)	An association with mortality is not an effect on mortality. A C–R association does not suggest that exposure poses a public health risk, unless the association is causal.
"Residential ambient air pollution exposures were *associated* with mortality … our study is the first to assess the *effects* of multiple air pollutants on mortality with fine control for occupation within workers from a single industry." Hart et al. (2011)	Associations with mortality are not effects on mortality.
"Each increase in PM2.5 (10 μg/m3) was *associated* with an adjusted increased risk of all-cause mortality (PM2.5 average on previous year) of 14% … These results suggest that further public policy efforts that reduce fine particulate matter air pollution are likely to have continuing *public health benefits*." Lepeule et al. (2012)	Associations do not suggest that public policy efforts that reduce exposure are likely to create public health benefits unless the associations reflect manipulative causation.
"Ground-level ozone (O3) and fine particulate matter (PM2.5) are *associated* with increased risk of mortality. We quantify the *burden* of modeled 2005 concentrations of O3 and PM2.5 on health in the United States. … Among populations aged 65–99, we estimate nearly 1.1 million *life years lost from PM2.5 exposure* … Among the 10 most populous counties, the percentage of deaths *attributable to PM2.5* and ozone ranges from 3.5% in San Jose to10% in Los Angeles. These results show that despite significant improvements in air quality in recent decades, recent levels of PM2.5 and ozone still pose a *nontrivial risk* to public health." Fann et al. (2012)	In the absence of manipulative causation, statistical associations between pollutant levels and mortality risks do not quantify effects caused by exposure on burden of disease or on life-years lost or on deaths, nor do they indicate a risk to public health.
"Ambient fine particulate matter (PM2.5) has a large and well-documented global *burden of disease*. Our analysis uses high-resolution (10 km, global-coverage) concentration data and cause-specific *integrated exposure-response (IER) functions developed for the Global Burden of Disease 2010* to assess how regional and global improvements in ambient air quality *could reduce attributable mortality from* PM2.5. Overall, an aggressive global program of PM2.5 mitigation in line with WHO interim guidelines *could avoid* 750 000 (23%) of the 3.2 million deaths per year currently (ca. 2010) attributable to ambient PM2.5." Apte et al. (2015)	The Global Burden of Disease IER functions are based on relative risk measures of association. They do not allow prediction or assessment of "how … improvements on ambient air quality could reduce attributable mortality" or avoid deaths unless the underlying relative risks represent (manipulative) causal relations.
"We use a high-resolution global atmospheric chemistry model combined with *epidemiological* concentration response *functions* to investigate *premature mortality attributable to* PM2.5 in adults ≥30 years and children <5 years. … [A]pplying worldwide the EU annual mean standard of 25 μg/m(3) for PM2.5 could *reduce global premature mortality* due to PM2.5 exposure by 17% … Our results reflect the need to adopt stricter limits for annual mean PM2.5 levels globally … *to substantially reduce premature mortality* in most of the world." Giannadaki et al. (2016)	Epidemiological exposure concentration-response associations and estimates of PM2.5-attributable mortalities based on them do not imply that reducing PM2.5 would reduce mortality, or allow such reductions to be predicted, unless the associations represent manipulative causal relations.
"*Relative risks* were derived from a previously developed exposure-response model. … Nationally, the *population attributable mortality fraction* of PM2.5 for the four disease causes was 18.6% (95% CI, 16.9–20.3%). … Aggressive and multisectorial intervention strategies are urgently needed to *bring down the impact* of air pollution on environment and health." Lo et al. (2016)	Relative risks and population attributable mortality fractions are measures of exposure-response associations. Such associations do not imply that interventions to reduce exposures would reduce risks of adverse responses unless there is a manipulative causal relation between them.

quite different interpretations of C–R relations. Such care has not been widely exercised in the literature reviewed next.

The importance of distinguishing between association and causation in interpreting C–R functions has recently been emphasized in comments on estimated human health benefits from reducing ambient concentrations of fine particulate matter (PM2.5) air pollution (Cox 2016; Frey 2016; McClellan 2016; Moolgavkar 2016; North 2016; Smith 2016). As detailed in Cox (2012), in studies that have been influential for estimating quantitative public health benefits attributed to PM2.5 reductions, the EPA has assumed a causal relation between PM2.5 reductions and reductions in both acute and chronic mortality for purposes of estimating health and economic benefits, while acknowledging in technical appendices that these causal assumptions are uncertain and are not clearly established by epidemiological data. This paper seeks to contribute to a sounder and less assumption-dependent scientific basis for understanding the probable human health consequences of changing pollution levels, by challenging the common practice of treating association as if it were causation for purposes of air pollution C–R modeling and by drawing attention to different concepts of causation and to less assumption-based methods for assessing causal impacts.

One specific purpose of this article is to critically review how C–R functions have been developed and used in public health risk assessments and risk management policy recommendations, focusing on fine particulate matter (PM2.5) and mortality risks as a prominent example. Because

interpretation of C–R functions for purposes of risk management decision-making is typically based on consideration of the wider context provided by multiple studies, rather than relying on analyses of any single data set, we focus on patterns of data analysis and causal interpretation that are prevalent across multiple studies. A second purpose of this article is to examine the extent to which associational C–R curves estimated from publicly available data permit accurate prediction of changes in *R* based on changes in *C*. To this end, we examine available data for the Boston and Los Angeles areas, recently identified in C–R studies as locations where further reducing PM2.5 concentrations is predicted to create substantial health benefits (Cromar et al. 2016; Schwartz et al. 2017). A third purpose of this paper is to examine how well C–R relations estimated for elderly people in Boston apply across the continent in Los Angeles. Finding the same C–R curve in such different geographic areas might suggest that it represents a predictively useful causal relation. Conversely, finding that different C–R curves hold in different locations or at different times in the same location would be evidence to the contrary.

The remainder of this paper is organized around these three purposes. The following section first examines how C–R functions have been developed and used in the epidemiological literature attributing adverse health effects to fine particulate matter. It critically discusses methods used in this literature to draw causal inferences and compares them to other methods for causal inference – what we call *information-based methods* – developed over the past century in genetics, engineering, economics and econometrics, statistics, systems biology, physics, computer science and artificial intelligence, machine learning, and other fields (Pearl 2009a). The next section applies these information-based causal inference methods to two publicly available data sets for PM2.5 and elderly mortality in the Northeast (Boston) and Southwest (Los Angeles air basin) and compares the results to each other and to results from associational (regression) modeling. Finally, we discuss limitations of the illustrative results presented and the information-based methods reviewed and acknowledge their rich and deep grounding and intellectual history. We conclude that current methods of information-based causal analysis used in other fields can provide a valuable complement to other techniques used in air pollution epidemiology. Applying them suggests that causal C–R functions, as they are usually described, understood, and applied, may not exist. Rather, published C–R functions quantify statistical associations that do not necessarily predict correctly how changing exposure concentrations would affect risk of adverse health responses. To better achieve this predictive goal, information-based methods can be used to address aspects of causality that are not well addressed by associational and potential outcomes methods.

Critical review and synthesis of literature on C–R relationships for PM2.5

Many papers have applied estimated C–R regression relationships to quantify human health risks associated with fine particulate matter (PM2.5) and to project human health benefits from further reducing ambient PM2.5 levels. Various criticisms of statistical uncertainties, limitations, and biases in C–R regression estimates have been offered (e.g. Young & Xia 2013; Krstić et al. 2016). The following paragraphs refer to selected papers, including recent ones and ones frequently cited in regulatory risk assessments, to illustrate the following key themes:

1. Associational and causal C–R functions have commonly been conflated in the literature, usually without explicit discussion or careful consideration of what causation means (Maldonado 2013).
2. There are many possible non-causal explanations for reported positive C–R associations. These are usually not systematically addressed using appropriate methods such as multiple-bias analysis (Greenland 2005).
3. Recent attempts to address causality for PM2.5 and mortality have relied on modeling assumptions of unknown validity, such as that no omitted confounders are present. They leave unaddressed other crucial assumptions, such as that individuals with different exposures are otherwise exchangeable, or that individual-level and population-level causation are consistent (Maldonado 2013).
4. Modern information-based algorithms for causal inference can be applied to PM2.5 health effects data, although few studies have yet done so.

Many important past papers equate associational and causal C–R relations

Current risk assessments, benefits assessments, and recommendations for revising standards for criteria pollutants build on a decades-old scientific literature that routinely makes the following three implicit assumptions:

a. Associational and causal C–R functions can be modeled by the same curve (Pope et al. 2002);
b. This C–R curve can be estimated quantitatively from relevant data via regression models or other associational methods such as odds ratios, relative risks, attributable risks, and burden of disease estimates, perhaps augmented with human judgments based on the Sir Austin Bradford Hill considerations (Hill 1965) or other weight-of-evidence considerations (Fedak et al. 2015) such as the strength, consistency, temporality, and biological plausibility of associations.
c. C–R regression coefficients or other associational measures estimated from one set of locations and times can be applied to exposure concentrations for other locations and times to estimate excess mortalities caused by air pollution and the potential human health benefits that would be caused by reducing it (e.g. Chen et al. 2013).

Table 1 gives examples of statements from articles that make one or more of these assumptions. Examples of associational and causal language are italicized (all emphases added). Brief comments in the right column note where

stated causal interpretations do not follow from the associations presented. These articles include many that have been cited by EPA and others in deliberations on PM2.5 risks and regulations in recent decades, continuing up to the present.

Such examples can be multiplied many-fold. Published papers on health effects attributed to pollutants typically move freely between associational and causal interpretations of C–R associations. They routinely present positive C–R associations as implying or suggesting that reducing air pollution would improve public health. But this is true only if the C–R associations are in fact causal – more specifically, only if they describe manipulative causality (Woodward 2013), as opposed to associational or counterfactual or predictive causation; these concepts are reviewed later. None of these papers establishes that the presented C–R associations do in fact describe (manipulative) causation, so none is suitable for predicting the effects on *R* of changing *C* by reducing air pollution. Of course, even if many published causal conclusions do not follow from their stated associational premises, this does not necessarily imply that either the premises or the conclusions are mistaken. It only implies that the methods used to draw causal conclusions are not reliable, They cannot be depended on to yield correct conclusions and may even yield contradictory results in the hands of different investigators (Glaeser 2006; Dominici et al. 2014), although they might also yield correct conclusions on some important occasions, as in the example of cigarette-smoking and lung cancer.

There are many potential non-causal explanations for positive C–R associations

If most positive C–R associations – or at least those satisfying commonly discussed weight-of-evidence considerations such as strength, consistency, specificity, temporality, and biological plausibility (Hill 1965; Höfler 2005; Fedak et al. 2015) – were in fact usually causal, then the distinction between associational and causal C–R relations might be of little more than academic interest. But there are many possible non-causal explanations for such C–R associations. Depending on the details of the study design used, non-causal C–R associations can arise for any of the following reasons.

- *Coincident historical trends*. Both *C* and *R* might be falling over time, but not because either causes the other. An example is a study of mortality risks before and after coal burning bans (Clancy et al. 2002) discussed later. Non-stationary time series (e.g. statistically independent random walks) often exhibit significant correlations in the absence of causation because they each independently tend to have trends over any interval of observation. To the extent that they are all correlated with time, they tend to be correlated with each other – the problem known as "spurious regression" (e.g. Yule 1926). "History," meaning trends or events that cause effects to change following interventions, but not because of the interventions, has been recognized and controlled for as one of the standard "threats to internal validity" of causal inferences in social statistics and quasi-experimental studies since the 1960s (Shadish et al. 2002). These ideas have had limited, but useful, impact on improving causal inference in health risk research (Slack & Draugalis 2001).
- *Omitted confounders*. Many data sets that identify positive C–R associations do not include potential socioeconomic confounders such as income, education, and occupation of those exposed, or societal factors such as stress and behavioral and societal factors (Valberg 2003), or time-varying confounders such as wind speed (Morabito et al. 2014; Urban & Kyselý 2014) or winter colds and flus. Such unmeasured variables might exert far larger effects on mortality than environmental exposures and differ systematically between more- and less-exposed areas.
- *Residual confounding*. This arises when a confounding variable is incompletely controlled for, e.g. by using discrete levels of a continuous confounding variable such as age or temperature, which allows for exposure and response variables to remain confounded *within* the discrete levels; or by using relatively stiff splines or inflexible linear parametric models to partially control for continuous variables such as recent daily temperatures. For example, Wang et al. (2016) note that "Many studies have reported the associations between long-term exposure to PM2.5 and increased risk of death. However, to our knowledge, none has used a causal modeling approach or controlled for long-term temperature exposure, and few have used a general population sample." To address these gaps, they explain that "We *estimated the causal effects* of long-term PM2.5 exposure on mortality and tested the effect modifications by seasonal temperatures, census tract–level socioeconomic variables, and county-level health conditions… Specifically, we *estimated the association* between long-term exposure to PM2.5 and mortality while *controlling for geographical differences using dummy variables* for each census tract in New Jersey, a state-wide time trend *using dummy variables for each year* from 2004 to 2009, and *mean summer and winter temperatures for each tract in each year*" (emphases added). In addition to the usual problem of conflating "we estimated the causal effects" with "we estimated the association," using a single dummy variable for each year and mean summer and winter temperatures leaves room for substantial residual confounding of exposure-response relations by time and by temperature variations within the summer and winter seasons.
- *Modeling biases*, including biases from omitted variables, omitted error terms or classification error probabilities for predictors, biases in data selection or coding, model form selection biases, and model functional form specification errors, provide other well-known non-causal sources of association. They are sometimes addressed via "multiple-bias modeling" (Greenland 2005; Höfler et al. 2007). They can greatly affect conclusions. As illustrated in our example with Models 1 and 2, different choices about which predictors to include on the right side of a regression model can change the sign, as well as the magnitude, of the C–R regression coefficient. Such findings have led some prominent researchers to conclude that "There is a growing consensus in economics, political science, statistics, and other fields that the associational or

regression approach to inferring causal relations – on the basis of adjustment with observable confounders – is unreliable in many settings" (Dominici et al. 2014). Sources of modeling bias can also interact in subtler ways. For example, suppose that mortality rate and exposure concentration depend only on temperature, but not on each other, via the following structural equations:

$$mortality\ rate = 100 - (0.1 \times temperature)^2$$
$$concentration = 100 - temperature,\ \text{for}\ 0 < temperature < 100.$$

- If a large data set is created by sampling temperature values from the interval from 0° to 100°, with each observed value of temperature and concentration differing from its true value by a small random error independently and identically uniformly distributed between −2 and 2, then fitting a linear regression model for mortality rate as a function of observed concentration and temperature will show that mortality rate is significantly *associated* only with concentration, and not with temperature, even though by construction mortality rate *depends* only on temperature, and not on concentration. Fitting a misspecified (linear) regression model yields a statistically significant C–R regression coefficient in the absence of a causal relation. Of course, in this simple example, regression diagnostics (e.g. plotting the data and residuals) would reveal the need for a nonlinear (quadratic) term rather than a linear term for temperature, allowing the incorrect conclusions to be avoided. But for the large and complex epidemiological models commonly used in practice, regression diagnostics and multiple-bias analyses that would allow effects of model specification errors and omitted errors-in-variables to be corrected are often not presented, potentially allowing modeling biases to affect results in unquantified ways (Greenland 2005).

Some of these possible non-causal explanations are sometimes mentioned in papers reporting positive C–R associations. However, they are seldom systematically listed and refuted by data, leaving readers uncertain about whether reported associations reflect reliable truths about the world or are only artifacts of modeling choices (Greenland 2005; Glaeser 2006; Dominici et al. 2014).

That potential non-causal explanations for C–R associations occur commonly in practice makes it important to report model diagnostics and to address such rival explanations before interpreting reported C–R associations as having causal significance or policy relevance (Greenland 2005).Yet, influential papers such as those in Table 1 often do little more than argue that smoking or selection biases are unlikely to explain the full C–R association before presenting a causal interpretation and urging policy-relevant recommendations based on it.

Recent papers draw causal conclusions from observational data by making unverified assumptions

Encouragingly, the more recent literature has started to address the question of causation more explicitly. This is most often done by introducing unverified modeling assumptions to justify interpreting regression coefficients, associations, and differences of means or proportions as if they were measures of causal impacts. The following types of studies and modeling assumptions have been used to support important causal claims about C–R associations for PM2.5 and mortality.

- *Intervention studies*, such as a widely cited study of effects on mortality risks of coal burning bans in Dublin County, Ireland (Clancy et al. 2002), assume that if health risks change *following* an intervention, then the change is *caused by* the intervention. This assumption was tested a decade after the original Dublin study in an updated study that compared mortality rates in areas affected and not affected by the bans (Dockery et al. 2013). In contrast to the original study, which had been used to justify policy decisions to extend coal-burning bans in Ireland based on a much-publicized belief that cleaner air had been found to cause reduced mortality, the updated study using control groups concluded that the bans had produced no detectable reductions in total or cardiovascular mortality rates (Dockery et al. 2013). As explained by Zigler and Dominici (2014), "However, even when studying an abrupt action, threats to causal validity can arise, as illustrated in extended analyses of the Dublin coal ban that revealed that long-term trends in cardiovascular health spanning implementation of the ban – not the coal ban itself – contributed to apparent effects on cardiovascular mortality." Since the 1960s, the "one-group pretest-posttest design" used in the original study has been identified by social statisticians as inappropriate for causal inferences, since it leaves uncontrolled the threat of coincident historical change, as well as other threats to valid causal inference (Campbell & Stanley 1963, p. 7). By contrast, a pretest–posttest control group design is appropriate for causal inference (*ibid*; Rich 2017). It can show that a large reduction in particulate pollution had no detectable effect on total mortality, as in Dublin, if that is the case; or it can provide strong evidence that high pollution levels cause excess mortalities if mortality rate spikes where and when air pollution spikes – such as in Donora in 1948, or London in 1952 – but not otherwise, e.g. in the same cities and months a year earlier or later than the high pollution episode, or in other cities with similar temperatures, humidity, influenza rates, etc. but without the high air pollution.
- *Instrumental variable* (IV) studies assume that a variable (called an "instrument") directly affects exposure but not response and is itself unaffected by unmeasured confounders (Baiocchi et al. 2014). Variation in response due to the variation of the instrument can then be used to estimate the effect on responses of the changes in exposure associated with variations in the instrument, without any effects from unmeasured confounders. For example, Schwartz et al. (2017) noted that "While many time series studies have established associations of daily pollution variations with daily deaths," most of these associations have been found at relatively high exposure

concentrations, and "causal modeling approaches are also lacking." To address this lack, they developed an IV approach that "combined height of the planetary boundary layer and wind speed, which impact concentrations of local emissions, to develop the instrument for PM2.5, BC, or NO2 variations that were independent of year, month, and temperature." They "conclude that there is a causal association of local air pollution with daily deaths at concentrations below EPA standards. The estimated attributable risk in Boston exceeded 1800 deaths during the study period, indicating important public health benefits can follow from further control efforts." However, the conclusions of such IV analyses depend on the validity of the assumptions that the instrument affects exposure but not response, and that it is unaffected by omitted confounders. Whether these assumptions are valid is usually unknown. As noted by O'Malley (2012), "IV analyses make strong assumptions that cannot be conclusively tested by the data" and different modeling choices, e.g. about how to treat lagged values, can yield very different results. Schwartz et al. (2017) assume that their instrument "is unlikely to be correlated with other causes of death," but the validity of this assumption is unknown, especially since wind speed, a component of the instrument, has recently been found to be correlated with cardiovascular mortality (Urban & Kyselý 2014), elderly mortality (Morabito et al. 2014), and stroke symptom onset (Kim et al. 2016).

- *Regression discontinuity designs* (RDDs) assume that if exposures and effects are significantly different for people with values of some characteristic (such as age) above *vs.* below an arbitrary threshold (such as an age threshold for legal drinking), then the difference in effects is caused by the difference in exposures, rather than by other differences. This assumption will be mistaken if the difference in effects is instead caused by other differences between people above and below the threshold (as might happen if the legal drinking age is also the legal age threshold for other activities such as gambling, smoking, driving without supervision, employment in certain occupations, and so forth). In air pollution health effects research, RDDs have been used to attribute differences in life expectancies between people living north and south of a river to differences in air pollution rather than to other regional differences in exercise or other variables, but this attribution remains an untested assumption (Chen et al. 2013). RDDs typically yield biased effects estimates unless the functional form of the C–R relationship is correctly specified.
- *Counterfactual* and *potential outcome* models, including *difference-in-differences* models, assume that differences between observed responses to observed exposure concentrations and unobserved model-predicted responses to "counterfactual" exposure concentrations are caused by the differences between the observed and counterfactual exposures, rather than by errors in the model or by systematic differences in other factors such as distributions of income, location, and age between the more- and less-exposed individuals. For example, Wang et al. (2016) assume that death count (the potential outcome) depends on predictors via the parametric model

$$\ln\left(E\left(Y^{a}_{c,t}\right)\right) = \beta_0 + \beta_1 a + \beta_2 Z_c + \beta_3 U_t + \beta_4 W_{c,t} + \ln(P_c)$$

where

- $Y^{a}_{c,t}$ = number of deaths that would occur in the population of census tract c were it exposed to a in year t;
- Z_c represents spatial confounders that vary among census tracts but not over the time period of the study;
- U_t represents confounders that vary over time but not among census tracts;
- $W_{c,t}$ represents confounders that vary over time and among census tracts, and
- $\ln(P_c)$ is the natural log of the population in census tract c.

If the assumed additive form of the right side is correct, then the change in deaths from one year to the next within a census tract should depend only on time-varying terms, and the difference in these changes (i.e. the "difference in differences") in deaths across census tracts should depend only on factors such as exposure that differ over time and between census tracts. Assuming that these differences-in-differences are caused only by differences in exposures, the authors conclude that "Under the assumption of the difference-in-differences approach, we identified a causal effect of long-term PM2.5 exposure on mortality that was modified by seasonal temperatures and ecological socioeconomic status." But the validity of the model assumptions is unknown. Counterfactual causality is based on estimating what *would have* happened had exposure conditions been different. Since what would have happened is not observed, models are used to guess at what would have happened; if these guesses are wrong, then their causal conclusions may be wrong. In competitive evaluations (Hill 2016), potential outcomes methods exhibit 20-fold greater errors and biases than other methods of causal analysis, as discussed below, reflecting the fact that they are sensitive to models of unknown validity.

Commendably, such limitations have been meticulously noted in some recent papers advocating and applying potential outcome methods. For example, Zigler et al. (2012) state that "Our analysis of the CAAA estimated that 1991 nonattainment designations for PM10 did causally reduce Medicare mortality in 2001, and that there are important causal pathways through which this effect occurred without affecting average ambient concentrations of PM10 or O3 during 1999–2001. … Our results are predicated on the belief that after adjusting for demographic characteristics in 2000–2001 … and preregulation pollution levels, *there are no unmeasured factors relevant to air quality and mortality that differ systematically between attainment and nonattainment areas*. … Furthermore, *we used a relatively restrictive (exponential) spatial decay function* that was indexed by a single parameter, but more flexible (e.g. anisotropic) spatial covariance functions could provide better fit to pollution data and should be explored" (emphases added). More recently, Zigler et al. (2016) concluded on the basis of potential outcomes

modeling that "all-cause Medicare mortality and respiratory-related hospitalization rates were causally reduced in areas designated as nonattainment for PM10 during 1990–1995 compared with the rates that would have occurred without the designation," but then responsibly noted that "a key limitation of our analysis is the fact that we estimated the effect of the nonattainment designation by regarding all monitoring locations in a nonattainment area as 'treated' However, nonattainment designations … sometimes resulted in no action at all. … [In addition, the] prospect of unmeasured confounding remained a threat to the validity of our results." Such expositions deserve credit for highlighting the fact that the conclusions reached are not necessarily valid, insofar as they depend on assumptions that are not necessarily correct.

Crucial assumptions, such as that unmeasured confounding does not invalidate the conclusions, are usually left untested in the counterfactual/potential outcomes approach, and many of its ablest practitioners suggest that they are inherently untestable. Thus, for example, Petersen and van der Laan (2014) write that "In sum, the flexibility of a structural causal model allows us to avoid many (although not all) unsubstantiated assumptions and thus facilitates specification of a causal model that describes the true data-generating process. Alternative causal models differ in their assumptions about the nature of causality and make fewer untestable assumptions. … [A] formal causal framework can provide a tool for defining a statistical estimation problem that comes as close as possible to addressing the motivating scientific question, given the data and knowledge currently available, while remaining transparent regarding the additional assumptions required to endow the resulting estimate with a causal interpretation." Clearly stating the untested assumptions (called "convenience-based assumptions" by Petersen and van der Laan, and distinguished by them from "real knowledge") that have been used to draw causal conclusions from observational data is certainly an admirable part of the program of counterfactual and potential outcomes causal research. However, the net result is that causal inferences and effects estimates in this framework are no more certain than the untested assumptions that support them. In practice, the approach often ends up interpreting associations predicted from regression models as if they were causal, while making (and explicitly stating) the convenience-based assumptions needed to warrant such an interpretation. This leaves open the practical question of whether the assumptions and causal conclusions are valid.

- *Predictive causality methods* assume that if exposure helps to predict a response, then exposure might be a cause of response. For example, *Granger causality* between an exposure and a response time series (Kleinberg & Hripcsak 2011) relies on the idea that causes help to predict their effects. Technically, variable X is a Granger-cause of variable Y if the future of Y is not conditionally independent of the history of X, given the history of Y. Thus, nicotine-stained fingers can be a Granger cause of lung cancer, helping to predict it, even if cleaning one's fingers would have no effect on future lung cancer risk; a predictive cause need not be a manipulative cause. Granger causality does not protect against causal associations from omitted confounders, nor does predictive causality have any necessary implications for manipulative causality. These points are often ignored or misunderstood in recent epidemiology papers. For example, Schwartz et al. (2017) state that that "We also used Granger causality to assess whether omitted variable confounding existed … Granger causality … argues that omitted covariates that are correlated with time varying exposure and outcome are as likely to be correlated with tomorrow's exposure as yesterday's exposure." This is neither a correct description of Granger causality, nor a valid test for omitted confounders. For example, consider a time series in which cold snaps occur only rarely, and for a single day at a time; people then turn on their heaters and PM2.5 concentrations spike on the day of the cold snap and the following day; flu-like symptoms occur one day later (i.e. 2 days after the cold snap day); and finally both temperature and PM2.5 then return to their usual distribution of values. If PM2.5 and flu-like symptoms are measured and cold snaps are an omitted covariate, there would be no justification for assuming that cold snaps "are as likely to be correlated with tomorrow's exposure [i.e. PM2.5 on the day after flu-like symptoms are observed] as yesterday's exposure [i.e. PM2.5 on the day before flu-like symptoms are observed]." To the contrary, occurrence of cold snaps could be perfectly correlated with yesterday's spikes in PM2.5 and yet have no correlation with tomorrow's random values of PM2.5. Even if Granger causality had been correctly established, the Schwartz et al. (2017) conclusion "that there is a causal association of local air pollution with daily deaths at concentrations below EPA standards … indicating important public health benefits can follow from further control efforts" would not follow: finding a causal association based on Granger causality or IV logically has no necessary implications for effects of control efforts, because such causal associations do not necessarily reflect manipulative causality (Woodward 2013). Just as showing that nicotine-stained fingers help to predict lung cancer would not imply that cleaning fingers would reduce risk of lung cancer, so showing that polluted air is a Granger-cause of mortality would not imply that cleaning air would reduce risk of mortality.
- *Attributable risk* and *burden-of-disease* studies, as previously discussed, assume that if responses are greater among people with higher exposures, then this difference is caused by the difference in exposures, and could be reduced by reducing it. Typically, this assumption is made without any careful justification: it simply confuses association with causation. Examples are widespread, e.g. Fann et al. (2012), Lepeule et al. (2012), Schwartz et al. (2017), and Lo et al. (2016).
- *Computer-based modeling studies* assume that simulated impacts in a computer model predict real-world responses. Many recent studies simulate health impacts of changes in exposures by applying C–R impact factors or functions to simulated changes in exposure

distributions. Validation of such models for specific applications is crucial for determining their usefulness and trustworthiness, but in practice they are usually used without such validation. As one prominent example, the United States Environmental Protection Agency (EPA) provides a publicly available computer program, BenMAP, that quantifies the subjective judgments of selected experts about C–R relations. The technical documentation for BenMAP (US EPA 2015, Table E-1, Health Impact Functions for Particulate Matter and Long-Term Mortality, pages 60–61) explicitly and repeatedly states "no causality included" in summarizing health impact associations based on expert judgments. The text further explains that "Experts A, C, and J indicated that they included the likelihood of causality in their subjective distributions. However, the continuous parametric distributions specified were inconsistent with the causality likelihoods provided by these experts. Because there was no way to reconcile this, we chose to interpret the distributions of these experts as unconditional and ignore the additional information on the likelihood of causality." BenMAP's health functions for long-term impacts should not be interpreted causally without carefully addressing these caveats. Nonetheless, many investigators present the results of BenMAP calculations as if they yielded "Estimated Excess Morbidity and Mortality Caused by Air Pollution" (Cromar et al. 2016) despite the BenMAP documentation and even though BenMAP has not been validated as a causal model. Similar comments apply to the increasingly widespread use of statistical and computational models to infer changes in mortality from changes in exposure concentration based on previously estimated C–R slope factors, without any attempt to validate manipulative causation (e.g. Lin et al. 2016).

Although most current literature on formal causal modeling of PM2.5-associated health effects depends on unverified assumptions, several recent papers strongly emphasize their authors' subjective confidence and conviction that associations can and should be interpreted causally. For example, one recent paper asserts that, in Boston, "the association between PM2.5 and deaths is almost certainly causal," even while noting that the supporting analysis "relies on the untestable assumption of no unmeasured confounding" or related untestable assumptions (Schwartz et al. 2015). This leaves open the question of whether the untested assumptions are correct. It also ignores a substantial statistical literature on methods for testing the assumption of no unmeasured confounding (Marra et al. 2014).

Many papers that express confidence in causal interpretations of epidemiological associations cite animal and *in vitro* studies as providing additional evidence that supports the hypothesis of causality (e.g. Schwartz et al. 2017). Adducing such data, or more general discussions and conjectures about plausible biological mechanisms by which exposures might cause the effects attributed to them, as evidence that epidemiological associations should be interpreted causally is problematic for several reasons. As noted by McClellan (2016) in discussing one such paper, "much of this discussion is quite simplistic and, indeed, naïve with regard to the actual complexity of disease processes." For example, premises such as that exposure leads to increased levels of reactive oxygen species (ROS) in the lung and that elevated levels of ROS in the lung are found in certain lung diseases do not necessarily constitute valid evidence that exposure increases risk of lung diseases, as elevated ROS levels also occur in response to exposure in healthy people and animals and do not necessarily or usually indicate any pathological mechanism at work (Cox 2011). More fundamentally, disease processes take place *within individuals*, while epidemiological C–R functions, relative risk ratios, and regression coefficients describe associations at the level of *populations* of individuals. Discussion of causal mechanisms, factor interactions, and confounding at the individual level may have no clear implications for analogous phenomena at the population level unless the frequency distribution of heterogeneous individual types within the populations is well understood (Maldonado 2013). Yet, papers that adduce mechanistic evidence in the context of epidemiological data seldom address this crucial methodological point, implicitly assuming that human populations can be treated as if they were composed of homogeneous individuals, with mechanistic discussions at the individual level being simply scaled up to apply to populations.

A more sober assessment is that *all* the papers we have reviewed rely on implicit or explicit untested modeling assumptions – typically, one or more of those discussed above, such as that changes following an intervention were caused by it, that statistical or computer simulation models used to predict counterfactual outcomes do so correctly, that instrumental variables are valid, and that unmeasured confounders do not exist – to justify their claims about causation. All address statistical counterfactual or predictive causation rather than manipulative causation, although manipulative causation is what risk managers and policy makers need to be informed about to predict effects of proposed interventions such as regulations or coal burning bans. None of these papers demonstrates a manipulative causal C–R relationship between changes in exposure concentrations and changes in total mortality risks. Some papers refute previous statistical causal claims (e.g. Dockery et al. 2013). Thus, it appears that the literature on adverse public health effects that are solidly proved to be caused by PM2.5 exposure and to be preventable by reducing PM2.5 exposure is still in its infancy despite decades of association-based and assumption-based studies: sound studies that address manipulative causality are still very much needed.

Opportunities remain to apply modern causal inference algorithms

The literature reviewed on PM2.5 health effects can be summarized as follows.

1. Past key papers do not distinguish clearly between association and causation, or directly address manipulative causation. As pointed out by Wang et al. (2016): "Many studies have reported the associations between long-term exposure to PM2.5 and increased risk of death.

However, to our knowledge, none has used a causal modeling approach." Similarly, Schwartz et al. (2017) noted that "While many time series studies have established associations of daily pollution variations with daily deaths … causal modeling approaches are also lacking."

2. The previous section identified a few studies that seek to apply assumption-based causal modeling approaches such as intervention studies, instrumental variable models, counterfactual or potential outcomes models, predictive causation models, burden of disease models, and computer simulation models. However, such studies are rare compared to the large number that deal only with associations. All of them rely on untested modeling assumptions, of unknown validity, to justify interpreting associations as causation.
3. More importantly, none of the papers reviewed clearly distinguishes between statistical counterfactual or predictive causation and manipulative causation. None demonstrates that manipulating exposure concentrations would change future mortality risks, or that it has done so in the past (once statistical errors in causal interpretations are corrected, as in the Dublin coal burning ban studies discussed by Clancy et al. (2002) and Dockery et al. (2013)). Frequent claims that reducing exposure concentrations would reduce mortality appear to be based on a widespread misconception that one type of causation implies others, despite counterexamples such as nicotine-stained fingers being a Granger cause but not a manipulative cause of lung cancer.

In short, the preceding review finds that the very large literature making policy-relevant causal claims and predictions about the human health effects of reducing air pollution is *not* well supported by analyses appropriate for manipulative causation. Causal claims about the predicted health effects of reducing air pollution are typically derived by conflating causation with association and conflating manipulative causation with other types of causation.

Fortunately, an extensive technical literature provides algorithms for automated causal discovery, modeling and analytics that offer constructive means to address the preceding limitations of existing C–R functions. These algorithms have emerged from a confluence of research in economics and econometrics, neuroscience, bioinformatics and systems biology, physics, engineering, statistics and applied probability, philosophy and formal logic, computer science, machine learning, and artificial intelligence (Pearl 2009b). They clearly distinguish among manipulative, counterfactual, predictive, and other (e.g. exogeneity-based or dynamic transition-based) forms of causation and enable causal inferences that do not require untestable assumptions or subjective weight-of-evidence judgments (Pearl 2010).

This technical literature on algorithms and methods for causal inference from observational data is already large and is growing rapidly though articles published in sources such as *The Journal of Causal Inference*, *Artificial Intelligence*, *Neural Information Processing Workshops on Causality*, *Uncertainty in Artificial Intelligence* (UAI) conference proceedings, and documentation of algorithms implemented in R and Python. From comparisons of the competitive performance of dozens of causal inference algorithms on benchmark problems over the past decade, it is possible to distill a short guide to the principles and algorithms that generally work best in practice. As of 2016, most top-performing methods in current causal analytics competitions for observational data use some or all the following principles (Bontempi & Flauder 2015; Hill 2016).

- *Information principle*: Causes provide information that helps to predict their effects and that cannot be obtained from other variables.

This principle creates a bridge between well-developed computational statistical and machine learning methods for identifying informative variables that improve prediction of dependent variables such as health effects, and the needs of causal inference. It allows techniques of predictive analytics to be applied to screen variables for potential causation. In practice, the information principle is usually implemented via algorithms that identify *conditional independence* among variables (Pearl 2009a). By this criterion if effect Y is conditionally independent of exposure variable X, given the values of other variables, then X is not eligible to be a cause of Y. Granger-causality is the special case of this principle in which the future of Y must not be conditionally independent of the history of X, given the history of Y. Although conditional independence tests do not establish manipulative causation, they provide a useful screen for potential manipulative causes, insofar as they provide a condition that is usually *necessary* (but not *sufficient*) for manipulative causation. For example, discovering that nicotine-stained fingers are a Granger cause of lung cancer would allow nicotine-staining to be identified as a *potential* manipulative cause of lung cancer, but we know that smoking is the actual manipulative cause.

In practice, statistical dependencies among variables can be discovered and displayed by algorithms that generate Bayesian networks or classification and regression trees (CART models) (Young & Xia 2013) from data, such as the *bnlearn* or *party* packages package in R, respectively. A Bayesian network (BN) shows arrows between pairs of variables that are statistically dependent even after conditioning on all other variables. Arrows between variables are absent if they are conditionally independent of each other after conditioning on other variables (e.g. Pearl 2009b; Rottman & Hastie 2014). The structure of such networks can often be learned directly from large data sets, rather than having to be specified as a hypothesis or based on expert knowledge, by using algorithms that test for conditional independence and quantify conditional probability dependencies (Frey et al. 2003; Aliferis et al. 2010; *bnlearn* package documentation by M. Scutari, www.bnlearn.com/). A CART tree (or, more generally, a "recursive partitioning" tree) shows combinations of values of predictors (or of ranges of their values) that yield significantly different conditional distributions for the dependent variable, and in this sense provide useful information for predicting it. Such trees can be used to help learn Bayesian network structures and to quantify their

conditional probability tables from data (Frey et al. 2003; Aliferis et al. 2010).

- *Propagation of changes principle:* Changes in causes help to explain and predict changes in their effects (Wu et al. 2011; Friston et al. 2013). This applies the information principle to changes in variables over time. It can often be visualized in terms of changes propagating along links (representing statistical dependencies) in a Bayesian network or other network model (e.g. Friston et al. 2013). The goal of causal analysis in C–R modeling is to predict how changing exposure would change health effects, so studying how changes propagate among variables over time is of great interest. In practice, examining propagation of changes from exposure to response variables requires longitudinal data and analysis of information flows from lagged to present values of variables.
- *Nonparametric analyses*. Multivariate non-parametric methods, most commonly, classification and regression trees (CART) algorithms, are used to identify information dependencies among variables (e.g. Halliday et al. 2016). If no significant change occurs in the conditional empirical cumulative distribution function of a dependent variable as the value of an explanatory variable varies, for any combination of values of the remaining variables, then this lack of dependence does not support a conclusion that the explanatory variable is a cause of the dependent variable. The dependent variable is then *conditionally independent* of the explanatory variable, given the values of other variables. Effects are not conditionally independent of their direct causes. The useful fact that CART trees can also be used to test for conditional independence, with the dependent variable being conditionally independent of variables not in the tree, given the variables that are in it, at least as far as the tree-growing algorithm can discover, only became widely appreciated and applied in causal analysis and machine-learning after 2000 (e.g. Frey et al. 2003; Aliferis et al. 2010). CART trees can be automatically generated using freely available tree-growing recursive partitioning algorithms in R packages such as *party* or *rpart* or Python *scikit-learn*.
- *Model ensembles*. Rather than relying on any single statistical model, the top-performing causal analytics algorithms typically fit hundreds of nonparametric models to subsets of the data (e.g. CART trees grown on random subsets of predictors to help de-correlate their predictions) (Hernandez et al. 2015; Furqan & Siyal 2016). Averaging predictions of how the dependent variable depends on other variables over an ensemble of models usually yields estimates with lower bias and error variance than any single predictive model. Computational statistics packages such as the *randomForest* package in R automate construction, validation, and predictive analytics for such model ensembles and present results in simple graphical forms. For example, partial dependence plots show how predicted values of a dependent variable change as a single predictor is systematically varied leaving all other variables with their empirical distributions (http://scikit-learn.org/stable/auto_examples/ensemble/plot_partial_dependence.html). Applied to C–R functions, such partial dependence plots show how *R* varies with *C* when accounting for the effects on *R* of all other variables via an ensemble of CART trees. *If* this dependency represents manipulative causality, then the partial dependency plot indicates how the conditional expected value of *R* should be expected to change when *C* is manipulated, given the empirical joint distribution of other measured predictors on which R also depends. This is an important *if*. Even state-of-the-art causal inference algorithms typically do *not* allow manipulative causality to be inferred confidently in the absence of interventions. Thus, humility about what can be accomplished using only observational data is prudent. But Bayesian Networks, *randomForest* ensembles, and partial dependence plots do allow *possible* causal dependencies – meaning dependencies that satisfy the information principle – to be identified and quantified from data. They also have several important practical advantages, including

  - *Automating variable-selection and coding*, thus reducing opportunities for p-hacking and confirmation bias to affect the analysis and conclusions;
  - *Detecting and modeling high-order interactions* among variables. This can be done using trees representing conjunctions of predictor ranges that lead to significantly different conditional distributions of the dependent variable, with the depth of the tree corresponding to the order of interactions considered;
  - *Coping with model uncertainties* by using ensembles of models; and
  - *Avoiding model selection and misspecification biases* by using non-parametric methods to avoid having to make parametric regression modeling assumptions.

As noted by Hernandez et al. (2015), "Random Forest (RF) … is a popular method for dealing with high-dimensional data, mainly because of its computational speed and high accuracy. It is a non-parametric method and so does not make any major distributional assumptions about the data. RF automatically allows for non-linear interaction effects, a desirable property in many high-dimensional datasets … . However, as RF is a machine learning algorithm and does not use a statistical model it cannot provide probability-based uncertainty intervals as in a Bayesian setting." Hybrid algorithms such as Bayesian Additive Regression Trees using Bayesian Model Averaging (BART-BMA) have very recently been proposed as a promising approach to combine these strengths of RFs with the ability of Bayesian methods to generate posterior uncertainty intervals (Hernandez et al. 2015; Chipman & McCullogh 2016). Such hybrid methods improve on traditional Bayesian Model Averaging (BMA) for parametric regression models by using nonparametric trees and allowing the forms of dependencies, rather than just the predictors selected, to be varied and averaged over. However, although they may well turn out to represent a lasting and valuable advance, these comparatively recent Bayesian tree ensemble approaches are not yet as mature and well-tested as pure RF

ensemble methods, which will therefore be used for the computations in this paper.

The design principles for causal inference algorithms just described – the information principle, propagation of changes, use of nonparametric estimates, and averaging over ensembles of predictions – characterize *information-based approaches* to causal inference. They emphasize discovery of which variables in a data set are informative about, and hence help to predict, which others, without regard for statistical associations and without depending on any specific parametric modeling assumptions. Information-based approaches are very different from associational or assumption-based ones. For example, two variables can be highly informative about each other even if the statistical correlation between them is zero, as in the case of $Y = X^2$ where X is uniformly distributed between -1 and 1. Conversely, two variables Y and Z can be strongly correlated even if they are conditionally independent of each other, so that neither provides information that helps to predict the other, given the values of other variables. An example is the system $Y = X^2$ and $Z = X^2$, with the value of X determining the values of both Y and Z. In this case, Y and Z will be perfectly correlated with each other (but not with X), although neither causes the other and X causes both. Information-based methods complement the assumption in potential outcomes or counterfactual models, that differences in causes make effects *differ* from what they otherwise would have been, with the idea that differences in causes help to *predict* differences in their effects. This should be true even when using nonparametric and model ensemble methods to make the predictions, including testing for conditional independence and quantifying probabilistic dependencies. Whether one variable helps to predict another can be tested using observational data without making hypothetical modeling assumptions about what would have happened had exposures or other conditions been different from those observed, and without relying on specific parametric modeling assumptions of unknown validity, by using nonparametric methods such as Bayesian Network learning, tree-growing algorithms, and ensemble methods.

Empirically, causal inference algorithm performance and validation results from a recent causal inference competition (Hill 2016) yielded the following quantitative results for comparison of a tree-based algorithm to a counterfactual/potential outcomes algorithm (Inverse Probability of Treatment Weighting (IPTW)), averaged over 20 challenge data sets for which the correct data-generating processes were known to the competition organizers, but not to the causal inference algorithm designer competitors who submitted algorithms to estimate these known causal relationships.

- *Bias:* Nonparametric regression tree methods had a bias of -0.007 in estimating the causal effects from data, compared to a -0.15 bias, about 20-fold greater, for the IPTW counterfactual causality algorithm.
- *Error:* The nonparametric regression tree algorithm yielded a root mean-squared prediction error of 0.02 compared to 0.41 for the IPTW algorithm, i.e. again about a 20-fold difference.
- *Coverage probabilities and uncertainty interval lengths:* The nonparametric regression tree algorithm had smaller uncertainty intervals and larger coverage probabilities than the IPTW algorithm.

It thus appears that the nonparametric tree-based approach can complement counterfactual approaches in at least some cases, not only requiring fewer modeling assumptions, but also providing better performance.

A hands-on example: C–R modeling of annual changes in PM2.5 and elderly mortality rates in Boston and Los Angeles

This section moves beyond a critical review of the literature on C–R functions and causal inference algorithms by applying modern information-based causal inference principles and algorithms to publicly available data for Boston (Suffolk County, MA) and Los Angeles (South Coastal Air Quality Management District (SCAQMD), CA) area data. The goal is to explore what can be learned about associational and predictive causal C–R functions by these methods, recognizing that, despite their extensive development in the journals and literatures previously cited, information-based causal analysis methods have not yet been widely used or well vetted in air pollution health effects research and epidemiology. Accordingly, our exploration is cautious, seeking only to illustrate the insights produced by applying these algorithms to two example data sets, rather than to reach definitive general conclusions. However, the state-of-the-art in computational statistical software makes these methods easy to apply, and it is more instructive to do so than only to discuss the literature about them.

Data

We use two data sets to illustrate information-based causal analytics methods, one from the Boston area (Suffolk County, MA) for 2000–2013 and the other from California's SCAQMD, which contains Los Angeles, for 2007–2010. The data fields for both areas are similar. Table 2 shows the layout of the data for the SCAQMD, from Lopiano et al. (2015). (The full data set can be downloaded from http://cox-associates.com/downloads; it is data set "Sample1" in the CAT software

Table 2. Layout of data for PM2.5 concentration, weather, and elderly mortality ("*mortality75*") variables in the LA-SCAQMD.

Year	Month	Day	Mortality75	PM2.5	t_{min}	t_{max}	MAXRH
2007	1	1	151	38.4	36	72	68.8
2007	1	2	158	17.4	36	75	48.9
2007	1	3	139	19.9	44	75	61.3
2007	1	4	164	64.6	37	68	87.9
2007	1	5	136	6.1	40	61	47.5
2007	1	6	152	18.8	39	69	39
2007	1	7	160	19.1	41	76	40.9
2007	1	8	148	13.8	41	83	33.7
2007	1	9	188	14.6	41	84	37.5
2007	1	10	169	39.6	41	78	63.2
2007	1	11	160	19.2	37	66	85.9
2007	1	12	160	22.3	31	56	67.2
2007	1	13	166	11.7	27	55	40.4
2007	1	14	157	20.8	24	56	34

at that web site. The Boston data is "Sample4".) The full data set has 1,461 rows of data, one for each day from 1 January 1 2007 to 31 December 2010. The variables (columns) in Table 2, and their data sources, are as follows:

- Calendar variables *year*, *month*, and *day* identify when the data were collected. Each row of data represents one day of observations.
- *mortality75* is a count variable giving the number of deaths among people aged at least 75 dying on each day, as recorded by the California Department of Health at www.cdph.ca.gov/Pages/DEFAULT.aspx. (This variable was originally named *AllCause75*, but we renamed it as *mortality75*.)
- *PM2.5* is the daily average ambient concentration of fine particulate matter (PM2.5) in micrograms per cubic meter of air, as recorded by the California Air Resources Board (CARB) at www.arb.ca.gov/aqmis2/aqdselect.php.
- The three meteorological variables t_{min} = minimum daily temperature, t_{max} = maximum daily temperature, and *MAXRH* = maximum relative humidity, are from ORNL (http://cdiac.ornl.gov/ftp/ushcn_daily/) and the US Environmental Protection Agency (EPA) www3.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm.

Lopiano et al. (2015) and the above sources provide further details on these variables. For example, for the *Mortality75* variable, Lopiano et al. explain that elderly mortality counts consist of "The total number of deaths of individuals … 75 + years of age with group cause of death categorized as AllCauses … . Note accidental deaths were excluded from our analyses." The definitions of the populations covered and the death categories used are taken from the cited sources, but it is clear that average PM2.5 concentrations at monitor sites do not apply in detail to each individual, any more than the weather conditions describe each individual's exposure to temperature and humidity. Rather, these aggregate variables should be interpreted only as providing data from which we can study whether days with lower recorded PM2.5 levels, or lower recorded minimum temperatures, relative humidity, and so forth, also have lower mortality, and, if so, whether predictive causality or other causal relationships hold between them. In addition to these variables, we also include the derived variable *time*, defined as the total number of months since the start of the data set. Month and year are each treated as categorical variables, but time is a continuous variable that allows longer-term trends to be modeled.

The Boston data are very similar, but with *Dewpoint* replacing *MAXRH* as a measure of humidity, and years ranging from 2000 to 2013. PM2.5 data were obtained from the US EPA Air Quality System (AQS) website (www.epa.gov/ttn/airs/airsaqs/) for central-site monitoring locations in the Greater Boston Area; daily quality controlled local climatological data (QCLCD) were downloaded from NOAA; and individual-level mortality records were obtained from the Massachusetts Department of Public Health.

We focus on elderly people (≥75 years old) because past literature has pointed to the elderly as being especially susceptible to health harm from PM2.5 (Goto et al. 2016), with considerable conjecture about causal mechanisms such as PM2.5 increasing oxidative stress and cardiovascular disease risks (Wang et al. 2016). Past literature has also noted that "in a time-series study in Boston [moving] the time scale from days to months (i.e. 60 d) increased the estimated PM effect" (Laden et al. 2006, citing earlier work by Schwartz et al.). We therefore aggregate the PM2.5 and elderly mortality data to the monthly level, so that our C–R functions will describe relations between monthly averages of daily PM2.5 concentrations and monthly averages of daily mortality counts for people 75 years old or older. The key question for C–R modeling with these choices of endpoints and time scales is whether months with higher average PM2.5 have correspondingly higher elderly mortality rates after controlling for other factors. For studying how well changes in PM2.5 levels help to predict changes in elderly mortality rates, we follow Wang, Kloog et al. (2016) in using a change over a one-year time interval. The key research question is how well differences in PM2.5 from one month to the same month a year later in a given location (Boston or the SCAQMD) predict corresponding differences in elderly mortality counts.

To give a visual impression of the data, Figure 1 plots several of the variables for the 48 months from January 2007 (denoted as 200701) to December 2010. Monthly averages of daily minimum and maximum temperatures, PM2.5 concentrations, and mortality counts are shown. Since the SCAQMD has a population close to 20 times larger than Boston's, the Boston mortality counts are multiplied by 20 in Figure 1 to put them on the same scale. It is clear from visual inspection of Figure 1 that elderly mortality decreases as temperature increases, but it is much less clear whether PM2.5 variations help to predict elderly mortality variations. Answering that question is a challenge for C–R analysis.

Methods and analytic plan

Our analytic plan is as follows.

1. *Identify conditional independence and dependencies* among variables in Table 2 using nonparametric Bayesian network learning algorithms and regression trees. This step screens for possible causal relations using the previously discussed information principle that variable *X* is a potential cause of variable *Y* only if *X* provides information that helps to predict *Y* and that cannot be obtained from other sources.
2. *Repeat step 1 for changes in variables over a time interval of one year.* This implements the propagation-of-changes principle that changes in causes should help to predict changes in their effects.
3. *Quantify the C–R dependence (if any) between PM2.5 concentration and elderly mortality*, without further assessing whether it is causal, using a partial dependence plot generated by the *randomForest* package in R. As discussed in the previous section, this algorithm averages the results of hundreds of nonparametric regression trees to estimate how daily mortality count varies as PM2.5 is

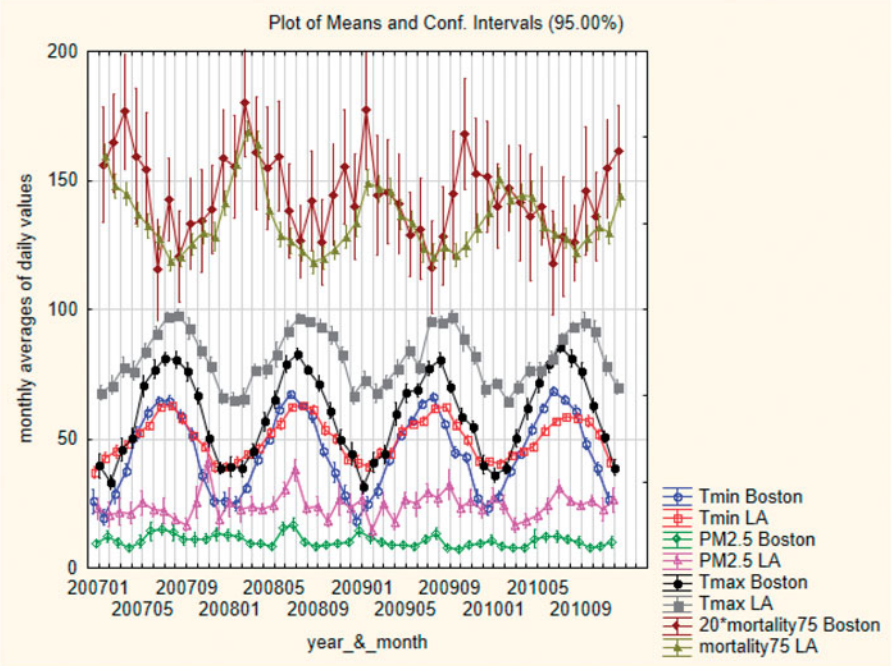

Figure 1. Monthly time series of selected variables, 2007–2010.

swept over its full range of values. The *randomForest* package documentation contains details of this process (https://cran.r-project.org/web/packages/randomForest/randomForest.pdf).

4. Compare the results from steps 1–3 to those from multiple linear regression analysis.

All analyses were performed using free computational statistics packages from the CRAN repository for the R project, https://cran.r-project.org/, as follows:

- Bayesian Network learning algorithms were run using the R package *bnlearn*, (www.bnlearn.com/) with all settings at their default values. This implements the information principle by using nonparametric machine learning algorithms to discover conditional independence and dependencies among variables.
- Regression trees were grown using the R package *party* (https://cran.r-project.org/web/packages/party/party.pdf) to further explore and visualize multivariate dependency and conditional independence relationships.
- *randomForest* model ensembles were generated by the R package *randomForest*, (https://cran.r-project.org/web/packages/randomForest/randomForest.pdf) to quantify associations between two variables, controlling for the levels of other variables using multiple nonparametric models.
- Additional R packages (*car*, *MASS*, *leaps*, *MSBVAR*) were used for parametric regression analyses and Granger causality testing.

To facilitate easy replication and interpretation by investigators not familiar with these packages, we accessed all R packages and displayed results using the Causal Analysis Toolkit (CAT), a free add-in for Microsoft Excel developed by the author and George Washington University Regulatory Studies Center for assessing the causal impacts of regulations using R packages (GWU 2016). The analyses can be replicated by clicking on the "Analyze" button on CAT. Details on CAT are provided in an appendix to Cox (2016) and in a User's Guide at www.cox-associates.com/downloads/. Details of all algorithms and supporting statistical theory are given in the online documentation for the corresponding R packages.

Results and discussion

This section first discusses results for Boston and then compares them to results for the LA (SCAQMD) air basin. As a point of departure, we note that if the goal were simply to find associational models in which PM2.5 is significantly positively associated with elderly mortality after controlling for selected other variables via regression, it could easily be accomplished. Table 3 shows an example of such a regression model for the Boston data, with a highly significant regression coefficient ($p = 1.7E{-}9$) of 0.17 for PM2.5, implying that each 10 $\mu g/m^3$ increase in PM2.5 is associated with an average of 1.7 extra deaths per day, about a 22% increase, since the mean death count is 7.675 deaths per day. Of course, as discussed earlier, such associations do not necessarily reveal anything about causation, since the association could be explained by non-causal factors such as omission of time as a potential confounder, in this example. Once time is included, so that the downward secular trends in PM2.5 and mortality as functions of time (elapsed months) can be modeled, PM2.5 no longer appears as a statistically significant predictor of elderly mortality in multiple linear regression ($p = 0.25$ instead of 1.7E−9), indicating that it appears as a highly statistically significant predictor in Table 3 only because it is acting there as a surrogate for the omitted variable *time*. Applying the Granger causality test using the granger.test function in the R

Table 3. Example of a multiple linear regression model for Boston with a significant positive C–R coefficient for the association between PM2.5 and elderly mortality.

CAT_linear (mortality75, PM2.5, t_{max}, t_{min}, Dewpoint, month)
Dependent variable: mortality75

Residuals:

Min	1Q	Median	3Q	Max
−1.4800	−0.4655	−0.0493	0.4449	1.9475

Coefficients:

	Estimate	SE	*t* Value	Pr(>|*t*|)
(Intercept)	8.37718	0. 77798	10. 77	< 2e−16***
PM2. 5	0.17307	0.02696	6.42	1.7e−09***
t_{max}	−0.01625	0.03747	−0.43	0.665
t_{min}	−0.05547	0.06282	−0.88	0.379
Dewpoint	0.00464	0.03384	0.14	0.891
Month10	1.13162	0.59401	1.91	0.059****
Month11	0.47998	0.42894	1.12	0.265
Month12	0.01960	0.29656	0.07	0.947
Month2	−0.07521	0.27538	−0.27	0.785
Month3	0.33780	0.33338	1.01	0.313
Month4	0.91830	0.47917	1.92	0.057*****
Month5	0.88112	0.64737	1.36	0.176
Month6	0.56257	0.83408	0.67	0.501
Month7	0.68506	0.96497	0.71	0.479
Month8	0.98317	0.95095	1.03	0.303
Month9	1.11129	0.80365	1.38	0.169

Residual standard error: 0.712 on 152 degrees of freedom.
Multiple R-squared: 0.596; Adjusted R-squared: 0.556.
F-statistic: 14.9 on 15 and 152 degrees of freedom, *p*-value <2e − 16.
Significant codes: ***0.001; ****0.1; and the rest at 1.

package *MSBVAR* shows that PM2.5 and elderly mortality are Granger-causes of each other, meaning that past and current values of each are significant predictors of the future values of the other, even after conditioning on its own past values and current value. Again, this indicates likely confounding by one or more other variables, such as time and/or temperature, that are correlated with both PM2.5 and elderly mortality, thereby making each informative about the other in bivariate analyses when the confounders are omitted. In both cases, statistical significance of an association does not imply substantive or causal significance.

Figure 2 shows the structure of a Bayesian network (BN) learned from the same Boston data by the R package *bnlearn*. In this network, an arrow between two variables shows that they are not conditionally independent of each other. Absence of arrows between two variables shows that they are conditionally independent of each other given the values of other variables. The directions of the arrows do *not* necessarily reflect causality. They simply indicate one way to decompose the joint probability distribution of all variables so that it can be computed from the marginal distributions of the input variables, those with only outward-pointing arrows, and conditional probabilities for the values of each other variable, given the values of the variables that point into it (Pearl 2009a). Nonetheless, the Bayesian network is a useful guide to possible predictive causality, insofar as the information principle implies that one variable can be a cause of another only if they are adjacent (linked by an arrow) in the Bayesian network. Figure 2 indicates that the passage of time is informative about both elderly mortality and PM2.5 (both decline with time), providing a reason that they are Granger-causes of each other.

If the *bnlearn* algorithms for discovering conditional independence and dependencies were completely accurate and trustworthy oracles, we would be done: Figure 2 would imply that PM2.5 concentration is not a predictive cause of elderly mortality risk on the chosen time scale of months, and thus is also unlikely to be a manipulative cause. The highly statistically significant ($p < 1.7E{-}9$) regression coefficient for PM2.5 as a predictor of elderly mortality in Table 3 would then be an example of a statistical effect but not a causal one. (In this case, fitting a linear model for elderly mortality counts introduces model specification errors that can be reduced by including PM2.5 or other month-dependent variables, thus making PM2.5 a useful statistical predictor even if it has no causal impact on mortality.) However, no algorithms for learning Bayesian network structures are perfectly accurate, and it is therefore worth continuing the analysis assuming that a C–R relationship might exist and simply be too weak to have been detected by the *bnlearn* algorithms.

Figure 3 shows a regression tree grown on the same Boston data as in Figure 2 and Table 3 using the *party* package in R. This tree identifies combinations of ranges of values for multiple variables that lead to significantly different conditional distributions for the elderly mortality (*mortality75*) dependent variable. The variables selected by the tree-growing algorithm as informative for predicting elderly mortality include *time* and t_{min}, in agreement with the Bayesian network in Figure 2, but also include PM2.5, unlike the Bayesian network. To read the regression tree in Figure 3, note that each shaded "leaf" node at the bottom of the tree shows the conditional mean value of the dependent variable, elderly mortality, given the ranges of values for the variables in the path leading to that leaf. For example, the right-most leaf node has an average elderly mortality count of 6.646 deaths per day for $n = 53$ months with $t_{min} > 48.677^{\circ}$ and *time* > 45 months. The intermediate nodes show the *p* values from *F* tests for rejecting the null hypothesis that the

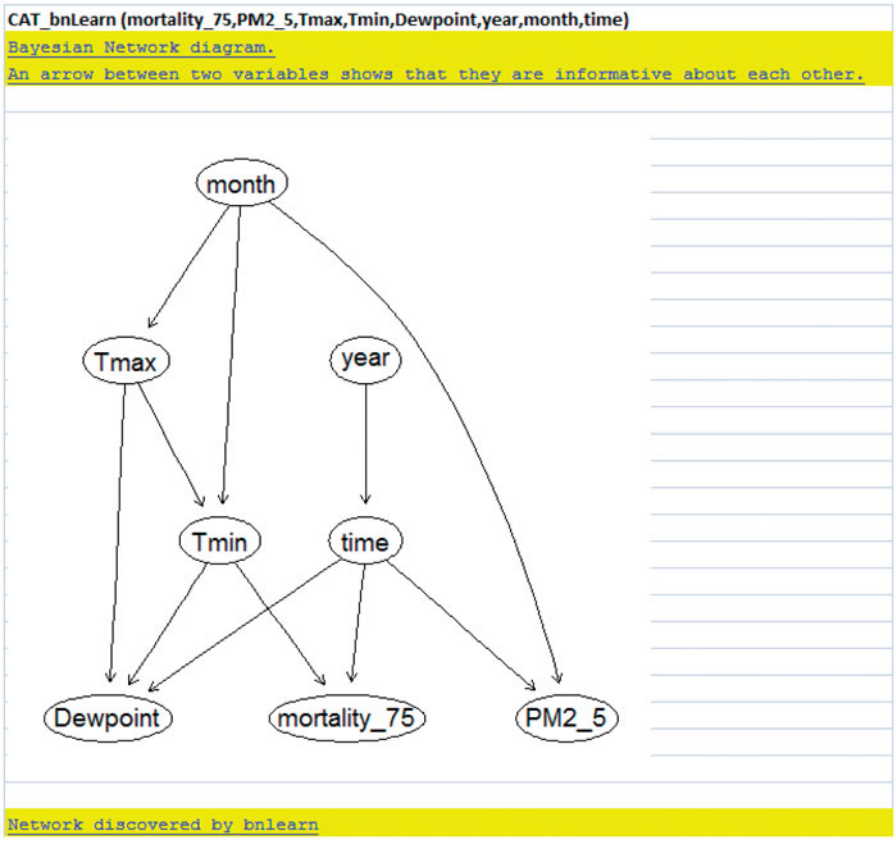

Figure 2. Structure of Bayesian network for Boston data discovered by R package *bnlearn*.

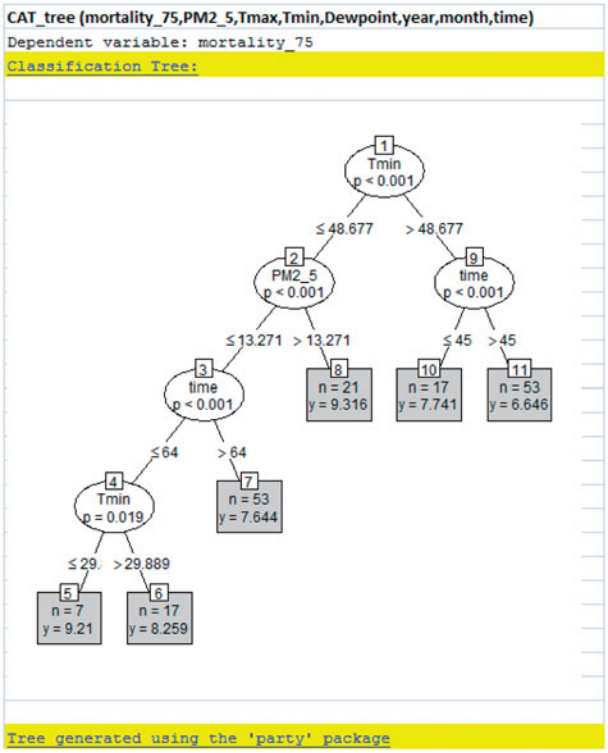

Figure 3. A regression tree for elderly mortality in the Boston data.

conditional distributions of elderly mortality are not different from each other on the left and right of each split. Figure 3 shows that months with warmer temperatures have lower average daily elderly mortality counts and that mortality rates have decreased over time. Relatively cold months with an average daily minimum temperature below 48.677° and average PM2.5 concentrations above 13.271 $\mu g/m^3$ have elevated elderly mortality.

Although individual trees are often not robust to perturbations in the data (e.g. growing trees on multiple random subsets of the data may produce many different trees), averaging results over ensembles of hundreds of trees for random subsets of the data using the *randomForest* R package yields the relatively robust C–R partial dependence plot in Figure 4. The same plot is shown on two different vertical scales, the left one emphasizing the variations that occur over their narrow range, and the right showing that the plot is almost flat when considered on a vertical scale that includes the origin. As average daily PM2.5 concentrations for specific months range over 2-fold, from $<10\,\mu g/m^3$ to $>20\,\mu g/m^3$, corresponding average daily elderly mortality counts range from about 7.64 to about 7.75 deaths per day, or roughly 1.01-fold. Although this variation could be due to omitted confounders or residual confounding within the intervals for continuous variables constructed by the tree-growing algorithm, no test performed so far rules out the possibility of a genuine predictive causal effect.

However, when the same analyses are repeated using *changes* in variables over a one-year period, the results are quite different. Only change in t_{min} is a significant predictor of change in elderly mortality, being adjacent to it in the Bayesian network for changes. Change in PM2.5 is also adjacent to change in t_{min}, so change in temperature could confound any association between change in PM2.5 and change

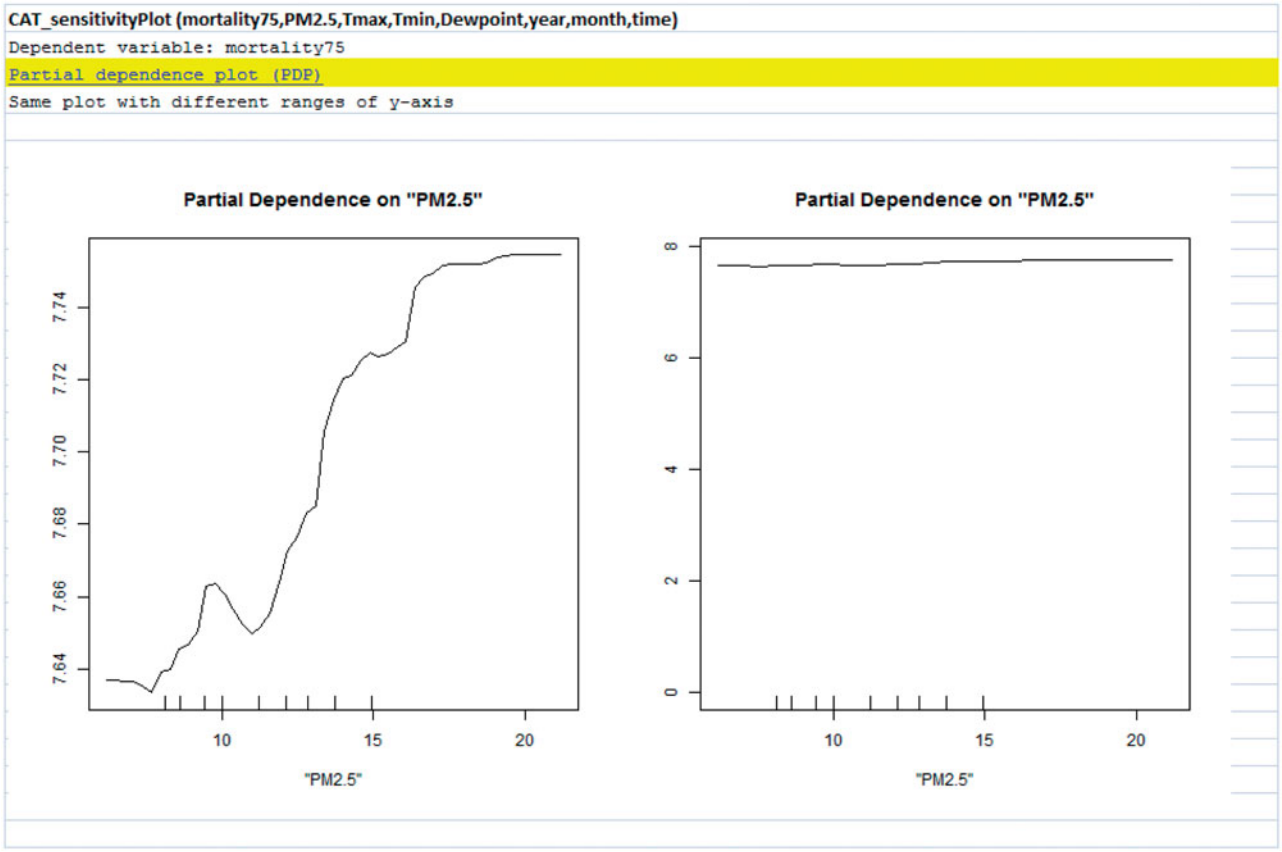

Figure 4. A C–R partial dependence plot for elderly mortality vs. PM2.5 in Boston.

in mortality. Regression trees do not identify any significant predictors of change in elderly mortality, but plotting the empirical cumulative distribution function (ECDF) of change in mortality (between a month and that same month a year later) conditioned on the upper and lower quartiles of values for change in t_{min} shows that increases of more than 2.45° in t_{min} are associated with reductions in elderly mortality risks (a leftward-shifted ECDF curve), while decreases in t_{min} by 2.19° or more are associated with increases in elderly mortality risks (rightward-shifted ECDF curves). By contrast, for PM2.5 the ECDFs for the top and bottom PM2.5 quartiles are not shifted in either direction but cross and re-cross each other. Thus, in this case, the Bayesian network algorithm proves more sensitive than regression trees in detecting what appears to be a genuine dependency relation between temperature and elderly mortality, and also perhaps more accurate than classification trees in identifying no significant dependency between PM2.5 and elderly mortality rate once other variables such as temperature and time have been accounted for by conditioning on their values.

Finally, we examine how well the C–R curve estimated in Figure 4 for Boston applies to the SCAQMD air basin. Figure 5 shows the main result. As average daily PM2.5 concentration in different months ranges from $<20\,\mu g/m^3$ to $>40\,\mu g/m^3$, average daily mortality counts stay essentially flat at 134.5 ± 0.5 deaths per day, with the slight residual C–R association between them being negative. PM2.5 and elderly mortality are adjacent in the Bayesian network for the SCAQMD, so the possibility of a slightly negative causal relation is not ruled out by these data. Ideally, confidence bands or uncertainty intervals would be provided with such plots, but, as noted by Hernandez et al. (2015), "However, as RF is a machine learning algorithm and does not use a statistical model it cannot provide probability-based uncertainty intervals as in a Bayesian setting." Instead, we simply note that the PM2.5-elderly mortality association is positive in Figure 4 and negative in Figure 5, so that both cannot be correct descriptions of one and the same causal relationship.

Comparing Figures 4 and 5, it is clear that estimated C–R functions cannot simply be transferred from one city or region to another. Table 4 confirms this for regression. In contrast to Table 3 for Boston, in a multiple linear regression model for the SCAQMD air basin, t_{min} has a significant negative association with elderly mortality for the LA region, but PM2.5 has no significant association with mortality. In a different regression model with only *PM2.5*, *MAXRH*, and *year* included as predictors, the C–R association between PM2.5 and elderly mortality is actually negative, −0.633, with $p < .10$, suggesting that confounding by t_{min} and *month* can affect the regression coefficient for PM2.5 if they are omitted from the model. Regression tree analysis identifies only t_{min} as a predictor of elderly mortality, with average daily mortality counts of 147.5 deaths per day when t_{min} is below 45.4°, compared to 120.6 deaths per day when t_{min} exceeds 58.1°. In the analysis of changes over a one-year period, however, neither regression tree analysis nor Bayesian network learning identified any dependencies among changes that helped to predict changes in elderly mortality, including changes in t_{min}, in contrast to Boston. Taken as a whole, these results suggest that the common practice of transferring C–R

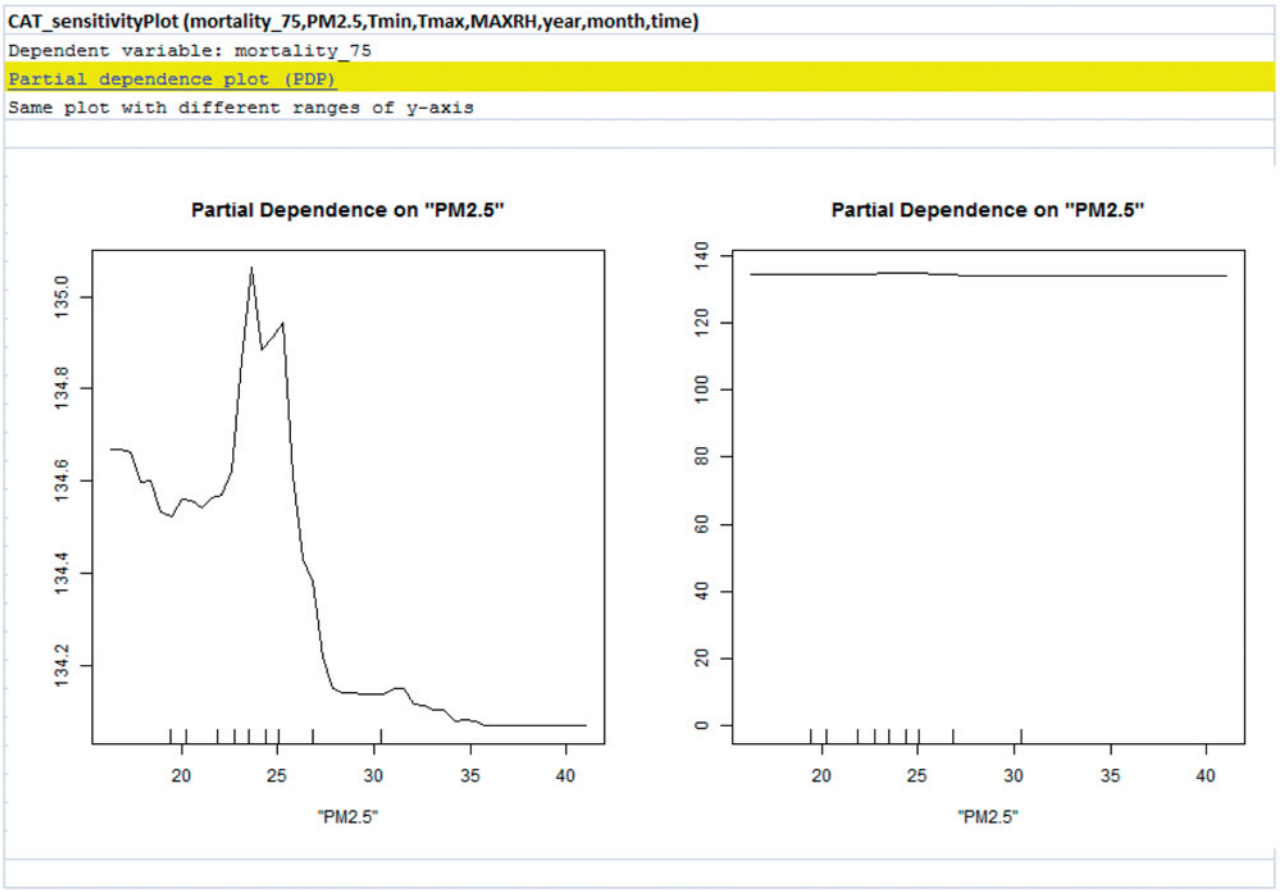

Figure 5. A C–R partial dependence plot for elderly mortality vs. PM2.5 in the SCAQMD air basin in southern California.

Table 4. Example of a multiple linear regression model for the SCAQMD air basin with a significant negative coefficient for the association between t_{min} and elderly mortality.

CAT_linear (mortality_75, PM2.5, t_{min}, MAXRH, month)
Dependent variable: mortality_75

Residuals:

Min	1Q	Median	3Q	Max
−8.044	−2.956	−0.296	1.778	16.436

Coefficients:

	Estimate	SE	*t* Value	Pr(>|*t*|)
(Intercept)	197.2808	20.6510	9.55	5e−11***
PM2 .5	0.1868	0.1956	0.96	0.3464
t_{min}	−1.1148	0.4988	−2.23	0.0323*
MAXRH	−0.0634	0.1623	−0.39	0.6985
month10	−9.1743	7.3792	−1.24	0.2225
month11	−14.2395	4.6642	−3.05	0.0045**
month12	−9.7250	4.0597	−2.40	0.0224*
month2	4.8563	4.3763	1.11	0.2752
month3	0.7252	4.4807	0.16	0.8724
month4	−7.3510	5.3307	−1.38	0.1772
month5	−9.0287	7.0615	−1.28	0.2100
month6	−6.3543	8.9425	−0.71	0.4823
month7	−8.0882	10.9337	−0.74	0.4671
month8	−7.0913	11.1635	−0.64	0.5297
month9	−6.8509	10.2753	−0.67	0.5096

Residual standard error: 5.21 on 33 degrees of freedom.
Multiple R-squared: 0.874; Adjusted R-squared: 0.82.
F-statistic: 16.3 on 14 and 33 degrees of freedom, *p*-value: 6.15e − 11.
Significant codes: ***0.001; **0.01; *0.05; and the rest at 1.

regression coefficients or associations estimated from data for one region to exposure changes estimated for a different area, as in the papers of Apte (2015), Cromar et al. (2016), Lo et al. (2016), and many others, is not well justified. Whether because of differences in PM2.5 composition or between populations or because associations in one area simply do not reflect stable causal laws useful for predicting impacts in another, it appears that there is no single C–R function that

correctly describes historical PM2.5-elderly mortality associations for both the Boston and the Los Angeles regions, controlling for the effects of other variables as in Figures 4 and 5. Nor does there appear to be a C–R function that allows changes in month-specific mortality risks to be predicted based on changes in month-specific PM2.5 concentrations, as no dependencies were found between these changes. These analyses do not support the usual assumption that a C–R function exists that can be used for these purposes.

Study uncertainties, limitations, and extensions

No proof of manipulative causality from observational data

The methods and conclusions illustrated in Figures 2–5 and our review of frameworks for causal inference and modeling of C–R functions have several limitations. First, the information-based approaches illustrated in Figures 2–5 only address whether conditions that are typically *necessary* for manipulative causation hold, such as whether exposures help to predict mortality rates after conditioning on observed confounders. These conditions do not provide *sufficient* conditions for inferring or quantifying manipulative causation based on observational data. They cannot, and are not intended to, "prove" manipulative causation based on observational data. Rather, they can be viewed as screening tests that establish whether, in the data analyzed, exposures are found to significantly predict health effects (and changes in exposure are found to significantly predict changes in responses), even after conditioning on other variables. If so, then it might be prudent to tentatively assume, for purposes of risk management, that C–R relations that have passed these screening tests are causal, while acknowledging that the tests stop short of definitively proving that exogenously changing (i.e. manipulating) future exposures would change future responses.

In the presence of unmeasured confounders, causality is difficult to establish for a relatively weak health risk factor such as ambient air pollution. However, consistency of the hypothesis of causation with available data can be established or refuted more easily. Even when confounders are not measured, their effects can often be detected as unexplained but significant associations between observed responses in disjoint subpopulation, such as between mortality rates among men over 75 and women under 75. Conditions for estimating effects of unmeasured ("latent") variables and for detecting causal paths despite unmeasured confounders have been developed (e.g. Spirtes et al. 1995). Focusing on necessary conditions that can be tested using available data helps to avoid the admitted philosophical and statistical difficulties of defining necessary and sufficient conditions for establishing manipulative causation, and also the need to make strong, untested assumptions to justify stronger causal conclusions. The price of this simplicity and testability is that the conclusions reached are only about predictive causation and not necessarily about manipulative causation.

No elucidation of causal pathways or explanatory causal mechanisms mediating C–R relationships

A second important limitation of the methods we have reviewed and illustrated is that they focus on whether there is evidence that a predictive causal C–R relation is present, and on quantifying it if so, but not on explaining how it works. This non-explanatory approach is sometimes referred to as "black box" causal analysis (Imai et al. 2011). Its goals are far more modest than those of automated causal discovery algorithms in general (Cooper et al. 2015; Spirtes & Zhang 2016). Current research in causal discovery seeks to provide algorithms and principles to support the common scientific ambition "to understand the mechanisms by which variables came to take on the values they have (i.e. to find a generative model), and to predict what the values of those variables would be if the naturally occurring mechanisms were subject to outside manipulations" (Spirtes 2010) – in other words, to address causal explanation, to reveal how causes produce their effects, and to make accurate predictions based on understanding of manipulative causation. This is part of the philosophical agenda *of causal realism*, succinctly described as follows (Lewis-Beck et al. 2003):

> There are two broad types of theories of causation: the Humean theory ("causation as regularities") and the causal realist theory ("causation as causal mechanism"). … Consider these various assertions about the statement, "X caused Y":
>
> - X is a necessary and/or sufficient condition of Y.
> - If X had not occurred, Y would not have occurred.
> - The conditional probability of Y given X is different from the absolute probability of Y ($P(Y|X) \neq P(Y)$).
> - X appears with a non-zero coefficient in a regression equation predicting the value of Y.
> - There is a causal mechanism leading from the occurrence of X to the occurrence of Y.
>
> The central insight of causal realism is that the final criterion is in fact the most fundamental. According to causal realism, the fact of the existence of underlying causal mechanisms linking X to Y accounts for each of the other criteria; the other criteria are symptoms of the fact that there is a causal pathway linking X to Y. … Causal realism insists, finally, that empirical evidence must be advanced to assess the credibility of the causal mechanism that is postulated between cause and effect. … A causal mechanism is a sequence of events or conditions, governed by lawlike regularities, leading from the *explanans* to the *explanandum*.

The challenge of elucidating causal pathways and processes describing *how* changes in exposures propagate through sequences of law-like biological, chemical, or other mechanisms to affect health, thus yielding explanations of C–R causal relations that can be used to predict accurately the health effects in individuals or populations of manipulating exposures, has been addressed via extensions of the principles and algorithmic approaches discussed in previous sections (Pearl 2014). Although a full discussion of these extensions is beyond the intended scope of this review, it is worth recognizing that (a) Detecting and quantifying causal C–R relations in populations stops well short of the mechanistic explanations addressed by advanced causal discovery

algorithms; but (b) The framework of information-based causal inference, including causal graphs (e.g. causal Bayesian Networks), conditional independence tests, and conditional dependence relationships quantified by partial dependence plots, can be extended to address explanation and description of causal pathways and mechanisms.

A traditional approach to elucidating causal pathways in epidemiology within a potential outcomes framework is causal *mediation analysis*. For example, Imai et al. (2011) present algorithms for estimating the "indirect effect" of an exposure or treatment variable on an outcome or response variable that is transmitted via an observed mediator, as well as for estimating the "direct effect" transmitted via other pathways not involving that mediator. They note that, as with other potential outcomes analyses, their estimation procedures require strong, unverifiable assumptions." As explained by Keele et al. (2015), "This gives causal mediation analysis the character of observational studies, where confounding between [the mediator and potential outcomes] must be ruled out 'on faith' to some extent." However, they make a constructive advance by presenting nonparametric estimation and sensitivity analysis algorithms, implemented in an R package, that quantify how large the errors introduced by unobserved pretreatment confounders might be. The problem of unobserved post-treatment confounders remains open, however.

Despite considerable effort and ingenuity invested in mediation analysis, many past applications are now widely understood to be incorrect or misleading due to failure to distinguish clearly between *conditioning* on values of variables and *manipulating* the values of those variables (Pearl 2014; Spirtes & Zhang 2016). For example, Pearl (2014) comments that much mediation analysis, including "principal stratification's mishandling of mediation" and other technical methods that he characterizes as products of "a century of blunders and confusions," stem from misguided efforts to use *conditioning* on a particular value of a mediator as a proxy for *holding* the value of a mediator fixed at that value. This confuses "seeing" with "doing," in Pearl's evocative terminology.

Even technically sophisticated algorithms developed and used for mediation analysis and estimation of causal impacts in recent decades, such as principal stratification analysis (Frangakis & Rubin 2002; VanderWeele 2011), and G-estimation of causal effects in the presence of time-varying confounding (Robins et al. 1992), do not overcome the fundamental problem that conditioning on observed values is quite distinct from setting those values (Pearl 2014). It is now widely appreciated that effects estimated from mean differences in responses conditioned on different values of an exposure need not coincide with the effects that would be caused by changing exposure from one value to another. To the contrary, conditioning or stratifying on some variables can create spurious C–R associations that have no causal interpretation. Algorithms for determining which subsets of variables to condition on to obtain unbiased estimates of direct causal effects of exposure concentration on response when causal DAGs are known have recently become available in special-purpose causal analytics software (Textor et al. 2011; Textor 2015). Conditioning or stratifying on variables without knowledge of the correct DAG model can be misleading for predicting how changing exposures would change outcomes; as previously noted, observing that heart attack risks are higher among elderly people with higher levels of aspirin consumption would not warrant any causal conclusions about how changing aspirin consumption would change heart attack risks.

Attempts to stratify individuals to more clearly reveal the effects that would be caused by changing exposures via pathways involving specified mediating variables do not solve the problem that conditioning is not manipulation. Like other data analysis strategies (such as matching, propensity scores, and instrumental variables) that try to approximate randomized assignments by conditioning on observed data, stratification methods do not address the key point that statistical relationships among observed levels of variables do not necessarily reveal how changing one would change others. Some experts have concluded that principal stratification "is of some use in assessing 'direct effects' [but] it is not the appropriate tool for assessing 'mediation.' There is nothing within the principal stratification framework that corresponds to a measure of an 'indirect' or 'mediated' effect" (Vanderweele 2011). Such assessments suggest that it is worthwhile to consider other approaches to elucidating causal pathways and mechanisms.

A complementary approach to causal explanation and prediction of effects of future interventions emphasizes the operation of sequences of causal mechanisms *within individuals* (Dash et al. 2013; Maldonado 2013). Causal discovery algorithms can be applied to data from cells, signaling pathways, biochemical and physiological processes, etc. within individuals (Faes et al. 2015; Schiatti et al. 2015; Lagani et al. 2016). Paying careful attention to the temporal sequences in which different observed variables respond to an exogenous change in the value of just one or a few of them can help to reveal the structure of causal networks and estimate the magnitudes and time delays in transmission of causal impacts among time series variables (Mohammad & Nishida 2011; Schiatti et al. 2015). Even without such exogenous shocks, quantifying predictive relations among variables using information-theoretic algorithms allows causal network structure and estimates of effect sizes and delays to be derived from time series data on multiple causally related variables under fairly general conditions (Sun et al. 2015). These approaches for elucidating causal mechanisms and pathways *within individuals*, developed largely in artificial intelligence (Dash et al. 2013) and systems biology (Lagani et al. 2016), complement the statistical strategies discussed and illustrated in previous sections for estimating differences in average responses across *subpopulations* of individuals with different exposures. The information-based algorithms on which we have focused can help identify which variables might cause which others and show the shape and magnitude of C–R relations in partial dependence plots, but additional algorithms can also help to quantify the timing of transmission of causal information among variables, adding further resolution to the description of causal processes (*ibid*).

Methods of causal inference that focus on explanatory mechanisms can also strongly complement black-box

statistical causal analysis by allowing valid causal conclusions to be drawn from studies of relatively few individuals if the studies suffice to elucidate the mechanisms of pathogenesis. Once it is understood how an exposure causes a response, this understanding can be used to make predictions and to generalize beyond the specific individuals and experimental or clinical conditions used to gain the initial understanding. For example, the discovery that smoking induces chronic unresolved inflammation in the lungs of susceptible smokers, followed by increased cell death and regenerative hyperplasia of the alveolar epithelium, provides a possible mechanistic basis for understanding (or, if it were not already known, predicting) that smoking will also increase risk of chronic obstructive lung disease (COPD) and lung cancer in these smokers (Cox 2011). This causal prediction could be arrived at without the need to study changes in COPD and lung cancer in large populations over time as smoking habits change. Similarly, suppose that it is known from toxicological and clinical data that inhalation of long, thin amphibole asbestos fibers increases repetitive injury of mesothelial tissue, resulting in (a) localized inflammation and release of TNF-α from alveolar macrophages attracted to the site of tissue injury and inflammation; and (b) upregulation of the expression of TNF-α receptors on mesothelial cells at the site, thus inducing mesothelial cells with DNA damage to survive and proliferate (via an NF-κB signaling pathway) instead of undergoing apoptosis and being safely removed (Yang et al. 2008). Such mechanistic knowledge, when available and correct, can be used to explain and predict health risks, such as the relatively high relative potency of amphibole fibers in causing malignant mesothelioma compared to the lower potency of cleavage fragments and mineral particles that do not have these biological activities (Bernstein et al. 2013). Conversely, mechanistic understanding can also help to identify when epidemiological associations are unlikely to be causal. For example, for PM2.5, Green et al. (2002) suggest that health effects *attributed* to PM2.5 exposure were unlikely on toxicological grounds to be *caused* by PM2.5 exposure, because most PM2.5 lacks the potency needed to induce such effects at realistic exposure concentration levels. Thus, mechanistic understanding can complement epidemiological data in showing how and why some exposures cause adverse effects while others do not.

Unresolved ambiguity and geographic heterogeneity of PM2.5 exposure metrics and unresolved negative studies

A familiar but important limitation of efforts to quantify C–R functions for PM2.5 is that "PM2.5" does not denote a well-defined, unique substance. Instead it refers to a heterogeneous collection of fine particulate matters. This raises the possibility that PM2.5 in one location might exert toxic effects while PM2.5 in a different location might not, simply because of differences in the composition of PM2.5 between the two locations. In this case, the concept of a causal C–R function that can be estimated from one or more data sets and then applied to estimate health benefits from reducing PM2.5 at other locations fails not because of such subtleties as that historical C–R associations are not valid predictive models, but simply because the same value of C is likely to have different meanings in different locations, measuring different things, and hence having different effects on R. Under such conditions, the existence of a single predictively useful C–R function for PM2.5 should not be expected.

There is empirical support for this concern, as well as a sub-literature that speculates about which specific components of PM2.5 might exert adverse health effects. For occupational exposures to a specific form of particulate matter, carbon black, Dell et al. (2006) applied standardized mortality ratio (SMRs) and Cox Proportional hazards regression modeling to quantify C–R association between lung cancer and respiratory disease mortality and cumulative inhalable carbon black exposure among over 6000 US carbon black workers. They found that no consistent C–R association between cumulative exposure to inhalable carbon black and respiratory disease mortality. On the public health side, Chay et al. (2003) "found that [1970 Clean Air Act] regulatory status is associated with large reductions in [total suspended particulates] pollution but has little association with reductions in either adult or elderly mortality" in the United States; the authors interpreted these negative findings cautiously, noting limitations in their study. Building on this work, Obenchain and Young (2017) concluded that reducing air pollution did not reduce deaths. Young and Xia (2013) found that there is geographic heterogeneity in air quality-mortality associations across the United States, with no effect of PM2.5 on life expectancy in the western United States. Young and Xia (2013) and Krstic (2013) both critically re-analyze previous reports associating PM2.5 with reduced life expectancy in the United States and highlight the importance of adequate control for potentially significant confounding factors and the need to consider influential outliers, specific variable-attributable effects, and geographical heterogeneity. Cox and Popken examined longitudinal data on PM2.5 and mortality rates in 100 US cities and concluded that there were many cities with statistically significant PM2.5-mortality associations, mostly (but not entirely) positive, but no clear evidence that PM2.5 is a Granger (predictive) cause of mortality. These studies looked for effects of PM2.5 reductions on mortality rates on time scales of years to decades. On a time scale of days to weeks, Zu et al. analyzed data from a natural experiment in which forest fires sent daily average PM2.5 concentrations above 60 micrograms per cubic meter in Boston and above 80 micrograms per cubic meter in New York for three days in 2002. They concluded that these spikes in PM2.5 did not produce "any discernible increase in daily mortality subsequent to the dramatic increase in ambient PM2.5 levels." On a wider spatial scale, an innovative study by Greven et al. (2011) used estimated "local coefficients" to try to reduce effects of unmeasured confounders in a very large spatio-temporal dataset (the Medicare Cohort Air Pollution Study, which included individual-level information on time of death and age on a population for over 18 million people between 2002 and 2006). They concluded that "Based on the local coefficient alone, we are not able to demonstrate any change in life expectancy for a reduction in $PM_{2.5}$." Outside the US, on a time scale of years to decades, as previously discussed, Dockery et al. (2013) reported that substantial (roughly

45–70%) reductions in black smoke in Ireland were accompanied by no detectable decreases in all-cause or cardiovascular mortality rates, correcting earlier reports to the contrary. Such negative findings clearly challenge the many assumption-based calculations, such as those in Table 1, that confidently predict that further reducing PM2.5 will substantially reduce all-cause mortality rates.

More generally, the fact that PM2.5 has fallen dramatically in many locations without detectably affecting mortality rates, contrary to projections from global burden of disease (GBD) and other models, invites explanation. One possible explanation is that C–R functions estimated by regressing past levels of mortality against past levels of PM2.5 exposure concentrations are not valid predictive causal models. Another possible explanation is that the composition of the PM2.5 exposures (or of the exposed populations) in locations such as Boston, New York, California and the western United States, Ireland, and other sites of negative studies may differ from the composition of PM2.5 implicitly assumed in models that project substantial mortality reduction benefits from further reducing PM2.5. Since the composition of PM2.5 to which these models apply is unspecified, their relevance to PM2.5 found in other specific locations is unknown. Without attempting to further resolve issues of geographic and compositional heterogeneity, and consequent limitations on the applicability of mortality reduction benefits calculations that assume that a single C–R function applies everywhere, it seems clear that such a universally applicable function might not exist simply because *C* lacks a unique, consistent definition in terms of causally relevant agents.

Generalizability of causal C–R functions across studies and applications

Apart from the challenges that arise from aggregating disparate substances under the heading "PM2.5," a more general issue for C–R functions is *generalizability* across studies – that is, the extent to which a C–R function estimated under one set of circumstances can be applied to others to correctly predict the impacts on responses of changing exposure concentrations. This problem has been studied since the 1960s in the quasi-experimental design literature on how to establish the *external validity* of causal inferences (Shadish et al. 2002). It has been addressed more recently within the information-based causal analytics framework of directed acyclic graphs (DAGs), conditional independence relationships, and conditional probability tables (e.g. Lee & Honavar 2013; Bareinbaum & Pearl 2013). As a simple example, consider a causal DAG model $X \rightarrow Y \leftarrow Z$ in which X is a manipulable decision variable such as exposure concentration, Y is a health variable of interest, such as mortality rate, and Z is a vector of covariates, such as age and sex. Suppose that a unique causal model specifying the probability distribution for Y values given X and Z values can be identified and estimated from some mix of observational and experimental or clinical data, yielding a conditional probability table (CPT) specifying the conditional probability of each possible value of Y for each set of input values for X and Z. This CPT can be denoted in symbols by $\Pr(Y = y \mid X$ is set to value x, Z has value $z)$ (or, more briefly, by $\Pr(y \mid \mathrm{do}(x), z)$), standing for the conditional probability that Y has value y given that X is set to value x and that Z has value z. (The use of "do" to distinguish variables whose values are set by a decision-maker from variables whose values are only observed is due to Pearl.) If this causal model is correct, and Y indeed depends only on X and Z, then the CPT can be transported to new settings in which the population distribution of covariate values is quite different from their distribution in the study or studies used to estimate the CPT. The CPT then represents a (probabilistic) causal law that should be invariant across settings, even though the joint frequency distribution of covariates, and hence the local C–R function for the average value of Y as a function of x, may differ in different populations.

In general, the C–R function $\mathrm{E}(Y \mid \mathrm{do}(x))$ will be different for different distributions of Z, implying that no single C–R function holds in different settings, simply because the C–R function estimated for one population reflects the distribution of covariates in that population as well as the causal law that determines risk from x for any given set of covariate values. Thus, contrary to widespread current practice, it would be misguided to believe that C–R effects estimated from one study can or should be applied unchanged elsewhere, or that effects of pollutants on human health estimated for a specific population and set of conditions in one country, such as China, "can be applied to other countries, time periods, and settings" (Chen et al. 2013). Yet, conditioning appropriately on Z provides so-called *transport formulas* for deriving the correct C–R function for a new population from the previously estimated CPT and the joint distribution of Z in the new population, if both are known. (Specifically, for our simple example, the probability that the mortality rate for a randomly selected individual will become y if the pollutant concentration is set to x is given by the probability identity: $\Pr(y \mid \mathrm{do}(x)) = \Sigma_z \Pr(y \mid \mathrm{do}(x), z)\Pr(z)$, where $\Pr(z)$ is the probability or relative frequency of covariates z in the target population for which causal impacts are to be estimated and $\Pr(y \mid \mathrm{do}(x), z)$ is the causal CPT, which may have been developed from previous studies elsewhere. The population C–R function for this target population is then given by the identity $\mathrm{E}(Y \mid \mathrm{do}(x)) = \Sigma_y \Pr(y \mid \mathrm{do}(x))^* y$.) Various important generalizations of this idea have been worked out in the recent epidemiological (Hernan & Vandeweele 2011; Schwartz et al. 2011) and computer science literatures, including methods for using causal relationships estimated from experiments and observations in several different source environments to estimate causal impacts of interventions from observational data for a target environment, when this is possible, and otherwise determining that it is not possible (Hernan & Vanderweele 2011; Bareinbaum & Pearl 2013; Lee & Honavar 2013). However, such transport formulas have not been developed and applied for causal C–R functions in the PM2.5 health effects literature to date, making it inappropriate to apply (i.e. transport) C–R functions estimated from studied populations to different populations, as encouraged by programs such as BenMAP (US EPA 2015). This appears to be an area where simply being aware of, and applying, modern causal analytics methods such as transportability conditions and formulas might be able to improve epidemiological and

public health risk assessments practice relatively quickly by enabling practitioners to adjust estimated causal functions to apply to different locations. On the other hand, to properly adjust for differences in distributions of response-related covariates across populations, considerable relevant biological knowledge may be necessary to identify the covariates that should enter the transport formulas, and enough data on the joint distributions of those covariates in the source and target populations must be available for the transport formulas to be applied.

Consideration of effects on different time scales

Our illustrative analyses used months as the time scale, essentially asking whether knowing average daily PM2.5 within a month helps to predict average daily mortality that month after conditioning on other variables such as temperature. Since daily values of the variable are available, the presence of effects on shorter time scales than months can also be examined. Cox (2016) reports results for the South Coastal Air Quality Management District based on analysis of daily data. Similar to the results for monthly data in the previous section, Bayesian Network, classification tree, and partial dependence plot analyses of daily again show that daily mortality rates depend on temperature but not on PM2.5. Including values of variables lagged by 1–7 days as predictors shows that today's elderly mortality depend not only on recent elderly mortality rates in the preceding 3 days (autocorrelation), but also on recent daily minimum temperatures (over at least the most recent 4 days). Elderly mortality does not depend directly on current or lagged PM2.5 values, but PM2.5 values are autocorrelated and lagged values of PM2.5, maximum relative humidity, and daily minimum temperature depend on each other (over a window of at least 4 days). These findings with daily resolution of the time scale confirm that elderly mortality is found to depend directly on daily minimum temperature but not directly on PM2.5 (same-day or lagged). The additional resolution from using daily data reveals transient autocorrelations and cross-correlations among these variables, and also relative humidity, over a period of a few days that are not apparent at the monthly level of aggregation, suggesting the importance of controlling for several days of temperature and humidity as confounders of the PM2.5-elderly mortality association. With time steps of months, these more detailed transients can be ignored. However, the existence of complex transients and interactions among variables on a time scale of days to weeks highlights the importance of carefully choosing time windows for case-crossover designs or dynamic regression models, since simply comparing exposure levels on nearby days with and without a death (e.g. Schwartz 2004) may misattribute to differences in daily exposure concentrations effects that are actually caused by lagged values of other variables that are correlated with such concentrations. For the Boston data, there are many days with missing data, especially in earlier years, so no analogous analyses are possible. Aggregating data to the monthly level resolves this problem, as the vast majority of days in each month have data from which monthly average values of all variables can be calculated. In addition, as previously noted, time steps of months may bring out more clearly than daily data the lasting effects of changes in pollution levels, temperature, or other variables on average daily mortality rates (Laden et al. 2006).

No discussion of the foundations and deep grounding of methods

Our review and application of information-based causal inference algorithms has deliberately emphasized principles and algorithms that have led to high performance in competitive benchmarking tests, while skipping over centuries of previous work. As noted by Pearl (2014), "Traditional statisticians fear that, without extensive reading of Aristotle, Kant and Hume, they are not well equipped to tackle the subject of causation, especially when it involves claims based on untested assumptions." Even the relatively short history of computational approaches to causal analysis of data, which is only about a century old, can be intimidating. Some of its key milestones are as follows:

- 1920s: *Path analysis* was introduced and developed by geneticist Wright (1921). This was the first approach to use directed acyclic graph (DAG) models in conjunction with quantitative analysis of statistical dependencies and independencies to clarify the distinction between correlation and causality. They have been so used ever since. Although Wright's path analysis was restricted to linear models, it can be seen as a forerunner of the Bayesian networks introduced some 70 years later, which generalize path coefficients to *conditional probability tables* (CPTs). These allow for non-parametric estimation of arbitrary (possibly non-linear) probabilistic dependencies among variables by specifying the conditional probabilities for the possible values of a variable, given each combination of values for the variables that point into it in a DAG model. (In practice, this conditional probability distribution or table at a node of a DAG model can be represented relatively efficiently as a classification tree for the node's value, given the values of its parents (inputs) in the DAG, rather than by explicitly listing all possible combinations of input values (Frey et al. 2003).) Path analysis and closely related linear structural equations models (SEMs) were extensively developed by social scientists and statisticians in the 1960s and 1970s and became a primary tools of causal analysis in the social sciences in those decades (Blalock 1967; Kenny 1979).
- 1950s: *Structural equations models (SEMs)* were developed as tools for causal analysis. For example, polymath and Nobel Laureate Herbert Simon defined causal ordering of variables in systems of structural equations (Simon 1953) and applied conditional independence and exogeneity criteria for distinguishing between direct and indirect effects and between causal and spurious correlations in econometrics and other fields (Simon 1954).
- 1960s: *Quasi-experiments* were introduced, standard threats to valid causal inference in observational studies were identified and listed, and statistical designs and

tests for overcoming them in observational studies were devised, most notably by social statisticians Campbell and Stanley (1963). These methods were extended and applied to evaluation of the success or failure of many social and educational interventions in the 1960s and 1970s, leading to a large body of techniques for program evaluation. The methods of data analysis and causal analysis developed for quasi-experiments, which consist largely of enumerating and refuting potential non-causal explanations for observed associations, have subsequently been extensively applied to "natural experiments" in which changes affect a subset of a population, allowing a quasi-experimental comparison of changes in responses in the affected subpopulation to contemporaneous changes in responses in the unaffected (control) subpopulation. These methods make it unnecessary to depend on already collected data on historical *C* and *R* levels, and instead allow potentially valid causal inferences about how changes in *C* change *R* (Rich 2017).

- 1965: *Hill considerations for causality introduced*. In 1965, Sir Austin Bradford Hill, expressing doubt that any valid algorithmic approach for causal discovery could exist, introduced a set of "considerations" to help humans make judgments about causality based on associations (Hill 1965). These considerations stand apart from the rest of the history of causal analysis methods, being neither greatly influenced by nor greatly influencing the technical developments that have led to successful current algorithms for causal discovery and inference. They have been enormously influential in epidemiology and regulatory risk assessment, however, where they have encouraged efforts to use judgment to interpret associations causally. Some attempts have been made to link Hill's considerations to counterfactual causality (Höfler 2005), but they play no role in current causal analysis algorithms, and the rates of false positives and false negative causal conclusions reached with their help have not been quantified. In Hill's judgment, "What they can do, with greater or less strength, is to help us to make up our minds on the fundamental question – is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect?" As a psychological aid to help epidemiologists, risk assessors and regulators to make up their minds, Hill's considerations have proved effective, but their performance as a guide for drawing factually correct conclusions about causality – especially manipulative causality – from observational data is less clear. Further discussion and synthesis of the Hill considerations with the information-based approaches to causality discussed earlier appears to be worthwhile, but is beyond the scope of this review.
- 1970s: *Conditional independence tests and predictive causality tests for time series* were developed to identify predictive causal relationships between time series, most notably by Nobel Laureate econometrician Clive Granger and colleague Christopher Sims, building on earlier ideas by mathematician and electrical engineer Wiener (1956). Granger (or Granger-Sims) tests for predictive causality have been extended to multiple time series and applied and generalized by neuroscientists analyzing observations of neural firing patterns in the brain (Friston et al. 2013; Wibral et al. 2013; Furqan & Siyal 2016).
- 1980s: *Counterfactual and potential outcomes techniques* were proposed for estimating average causal effects of treatments in populations, largely by statistician Donald B. Rubin and colleagues, building on work by statistician Jerzey Neyman in 1923. Over the course of four decades, specific computational methods put forward in this framework to quantify average causal effects in populations, usually by trying to use observations and assumptions to estimate what would have happened if treatments or exposures had been randomly assigned, have included matching on observed covariates (Rubin 1974), Bayesian inference (Rubin 1978), matching with propensity scores (Rosenbaum & Rubin 1983), potential outcomes models with instrumental variables (Angrist et al. 1996), principal stratification (Zhang & Rubin 2003), and mediation analysis (Rubin 2004). These methods have been influential in epidemiology, where they have been presented as suitable for estimating average effects caused by treatments or interventions. But they have also been criticized within the causal analysis community as being needlessly obscure, reliant on untestable assumptions, and prone to give biased, misleading, and paradoxical results in practice, in part because they do not necessarily estimate genuine (manipulative) causal effects (e.g. Pearl 2009a). From this perspective, the useful contributions of the potential outcomes framework can be subsumed into and clarified by methods of structural equations modeling (Pearl 2009b).

The 1980s also saw the introduction of classification and regression trees (CART) methods (Breiman et al. 1984). These would eventually provide nonparametric tests for conditional independence, useful for learning Bayesian network structures from data (Frey et al. 2003). They also provided the base nonparametric models for *randomForest* ensembles and related non-parametric ensemble algorithms now widely used in machine learning (Furqan & Siyal 2016).

- 1990s: *Probabilistic graphical models* were developed in great detail and given clear mathematical and conceptual foundations (Pearl 1993). These included Bayesian networks and causal graph models, together with inference algorithms for learning them from data and for using them to draw causal inferences and to estimate the sizes of effects caused by interventions. These methods are most prominently associated with the Turing Award-winning work of computer scientist Judea Pearl and his coauthors. They grew out of the intersection of artificial intelligence and statistics. They provide a synthesis and generalization of many earlier methods, including structural equations modeling (both linear and nonlinear), probabilistic causation, manipulative causation, predictive (e.g. Granger) causation, counterfactual and potential outcomes models, and directed acyclic graph (DAG) models, including path analysis. Conditional independence tests and quantification of conditional probabilistic

dependencies play key roles in this synthesis, as set forth in landmark books by Pearl (2000) and Koller and Friedman (2009). The full, careful development of probabilistic graphical models and algorithms created what appears to be a lasting revolution in representing, understanding, and reasoning about causality in a realistically uncertain world.

- 2000–Present: *Causal discovery and inference algorithms* for learning causal DAG models from data and for using them to draw causal inferences and to quantify or place bounds on the sizes of impacts caused by different interventions have been extensively developed, refined, tested, and compared over the past two decades. Important advances included clarifying which variables in a DAG model must and must not be conditioned on to obtain unbiased estimates of causal impacts in known DAG models (Shpitser & Pearl 2008; Textor 2015), as well as transportability formulas for applying causal relationships discovered and quantified in one or more learning settings to a different target setting (Hernan & Vanderweele 2011; Bareinbaum & Pearl 2013; Lee & Honavar 2013). Recent years have also seen substantial generalizations of earlier methods. For example, transfer entropy, a nonparametric generalization of Granger causality, quantifies the rates of directed information flows among time series variables. Introduced by physicist Schreiber (2000) and subsequently refined and extended by workers in computational finance and neuroscience (Wibral et al. 2013), transfer entropy and closely related methods appear to be promising for creating algorithms to discover causal DAG structures and quantitative dependency relationships and time lag characteristics from observations of multiple time series.

Even such an abridged list of milestones makes clear that causal analytics is now a large and deep field with a host of interrelated technical concepts and algorithms supported by a confluence of insights and methods from statistics, social statistics and program evaluation, electrical engineering, economics and econometrics, physics, computer science, computational finance, neuroscience, and other fields. Any brief survey must therefore be relatively superficial; full treatments run into thousands of pages (e.g. Koller & Friedman 2009), and even documentation for R packages implementing the key ideas can be hundreds of pages.

This deep grounding of current information-based causal analytics methods and algorithms, such as those in CAT, in nearly a century of computational methods backed by centuries of philosophizing about causality might well inspire a prudent humility (Pearl 2014). Yet, for the practitioner with limited time and a need to draw sound causal inferences from data, two relatively recent developments make even superficial understanding of key ideas and software packages highly useful. The first is that many formerly distinct causal analysis methods have now been synthesized and unified within the framework of information-theoretic methods and directed acyclic graphs. This framework brings together ideas from potential outcomes and counterfactual causation, predictive causality, DAG modeling, and manipulative causality (Pearl 2000, 2010). The second is the success of the object-oriented software paradigm in platforms such as R and Python. Modern software enables and encourages encapsulation of technical implementation details so that only key ideas and behaviors of software objects need be understood to use them correctly. This allows users with only a superficial understanding of exactly what a software package does and how it works to use it appropriately to do valuable tasks. For example, a user who understands only that causes must be informative about their effects, and that this can be indicated graphically by arrows between variables showing which ones are identified as being informative about each other and which are conditionally independent of each other, can use this limited understanding to interpret correctly the results of sophisticated algorithms such as those in the CAT package. As a practical matter, making tools such as Bayesian network learning algorithms, classification trees, and partial dependency plots widely available and easy to apply can complement insights from regression-based and other associational and counterfactual methods to reveal and quantify potential causal relationships in observational data.

Summary and conclusions

Our critical review of the literature on concentration-response (C–R) relations for fine particulate matter and mortality has identified the following challenges:

- C–R functions that describe historical associations do not necessarily predict how changing C would change R. This is partly because associations may not represent manipulative causal relationships, as when positive associations between baby aspirin consumption and heart attack risk, or between nicotine-stained fingers and subsequent risk of lung cancer, do not allow a valid prediction that reducing one would reduce the other.
- Almost all of the existing literature on PM2.5-mortality C–R functions deals with associations and not with causality (Wang et al. 2016).
- The few papers that do attempt to model causality in C–R relations for PM2.5 exposure and mortality fail to distinguish among counterfactual, predictive, and manipulative causality. Most of these papers follow a counterfactual approach that relies heavily on unverified modeling assumptions about unobserved potential outcomes. This amounts to assuming, rather than showing, that associations in regression models can be interpreted causally. Arguably, manipulative causation should be the gold standard for discussions of causality and the public health effects that would be caused by changing PM2.5 exposures. Although even state-of-the-art causal inference algorithms cannot definitively establish manipulative causality from observational data, they can identify several measures of predictive causality (Granger causality, information relations in DAG models, and so forth) that provide a valuable screen for evidence of potential manipulative causation.
- A large literature on modern causal inference algorithms for observational data has yet to be applied to C–R

modeling. The most successful methods in this literature, as assessed in competitive challenges, emphasize the information criterion (a necessary but not sufficient condition) that causes should be informative about their effects. Changes in causes should also be informative about changes in their effects. These principles can be applied to observed data using freely available R packages for predictive analytics. Nonparametric CART trees used to detect and quantify information about responses provided by exposure concentrations are among the most successful and most mature current algorithmic approaches to computational causal inference.

- Applying these methods to publicly available data from the Northeast (Boston) and Southern California (Los Angeles) shows that C–R associations found in Boston do not hold in the SCAQMD air basin. Quite significant (e.g. 2-fold) changes in average PM2.5 concentrations do not help to predict changes in average elderly mortality rates in either location, at least on the time scales of monthly averages separated by a one-year lag used in our example analyses.

Well-defined causal C–R functions do not necessarily exist for PM2.5 and elderly mortality, at least on the time scales considered. Different statistically significant associations hold in different areas, but they do not necessarily correspond to the theoretical construct of a C–R function that can be estimated at one location and applied to another to approximate how changes in concentrations would affect changes in public health. It seems highly desirable for future work to distinguish more clearly among statistical association, predictive causal, and manipulative causal C–R functions than past research has done. The fundamental premise that C–R functions exist that can predict the public health effects caused by reductions in pollutant concentrations needs to be carefully reexamined and tested, as it does not appear to hold in general.

Acknowledgements

The author acknowledges the enormously valuable comments received from six reviewers who were selected by the Editor. These comments prompted substantial expansions of the manuscript, additional references and discussions, and improvements in content and exposition.

Declaration of interest

The employment affiliation of the author is shown on the cover page. Cox Associates is a private firm providing research and consulting services, principally on risk analysis, analytics, operations research, and applied statistics issues, to private and public entities. The work described here was supported by Cox Associates and the American Petroleum Institute. Over the past five years, Cox Associates has received funding from the American Petroleum Institute (API) and the American Chemistry Council (ACC) and their members to analyze causal relations between exposure concentrations and adverse health responses for crystalline silica and PM2.5. The research questions asked, technical methods selected, and conclusions reached are solely those of the author. This paper benefitted from close proof-reading and copy-editing suggestions from API, but these reviews and suggestions were provided for the author's consideration without constraints that any of them be incorporated. The author has testified before Congress on matters related to data transparency and causation of adverse health effects by air pollutants. He is Editor-in-Chief of *Risk Analysis: An International Journal* and has contributed to and encouraged discussions of associations vs. causality for C–R functions in that journal. All of the views presented here are solely those of the author and in no way reflect any positions of the journal *Risk Analysis* or the Society for Risk Analysis, the API, the ACC or their member companies.

Data availability

The full data LA (South Coastal) data set analyzed in this paper can be downloaded from http://cox-associates.com/downloads; it is data set "Sample1" or "LA" in the CAT software at that web site, under the "Excel-to-R" button. The full Boston data is "Sample4" or "Boston".

References

Aliferis CE, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XS. 2010. Local causal and markov blanket induction for causal discovery and feature selection for classification Part I: algorithms and empirical evaluation. J Machine Learn Res. 11:171–234.

Angrist JD, Imbens GW, Rubin DB. 1996. Identification of causal effects using instrumental variables. J Am Stat Assoc. 91:444–455.

Apte JS, Marshall JD, Cohen AJ, Brauer M. 2015 Jul 7. Addressing global mortality from ambient PM2.5. Environ Sci Technol. 49:8057–8066.

Baiocchi M, Cheng J, Small DS. 2014. Instrumental variable methods for causal inference. Stat Med. 33:2297–2340.

Bareinboim E, Pearl J. 2013. Meta-transportability of causal effects: a formal approach. Paper presented at: Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, Scottsdale, AZ, USA.

Bernstein D, Dunnigan J, Hesterberg T, Brown R, Velasco JA, Barrera R, Hoskins J, Gibbs A. 2013. Health risk of chrysotile revisited. Crit Rev Toxicol. 43:154–183.

Blalock HM. 1967. Causal inferences in nonexperimental research. Chapel Hill (NC): UNC Press.

Bontempi G, Flauder M. 2015. From dependency to causality: a machine learning approach. J Machine Learn Res. 16:2437–2457.

Breiman L, Friedman J, Olshen R, Stone C. 1984. Classification and regression trees. Belmont (CA): Wadsworth.

Campbell DT, Stanley JC. 1963. Experimental and quasi-experimental designs for research. Boston (MA): Houghton Mifflin Company.

Chay K, Dobkin C, Greenstone M. 2003. The Clean Air Act of 1970 and adult mortality. J Risk Uncertainty. 27:279–300.

Chen Y, Ebenstein A, Greenstone M, Li H. 2013. Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy. Proc Natl Acad Sci USA. 110:12936–12941.

Chipman H, McCulloch R. 2016. Package 'BayesTree' [Internet]. [cited 2017 Mar 30]. Available from: https://cran.r-project.org/web/packages/BayesTree/BayesTree.pdf

Clancy L, Goodman P, Sinclair H, Dockery DW. 2002. Effect of air-pollution control on death rates in Dublin, Ireland: an intervention study. Lancet. 360:1210–1214.

Cooper GF, Bahar I, Becich MJ, Benos PV, Berg J, Espino JU, Glymour C, Jacobson RC, Kienholz M, Lee AV; the Center for Causal Discovery team, et al. 2015. The center for causal discovery of biomedical knowledge from big data. J Am Med Inform Assoc. 22:1132–1136.

Cox LA Jr. 2011. A causal model of chronic obstructive pulmonary disease (COPD) risk. Risk Anal. 31:38–62.

Cox LA Jr. 2012 May. Reassessing the human health benefits from cleaner air. Risk Anal. 32:816–829.

Cox LA Jr. 2016. Rethinking the meaning of concentration–response functions and the estimated burden of adverse health effects attributed to exposure concentrations. Risk Anal. 36:1770–1779.

Cox LA Jr. 2016. Rethinking the meaning of concentration-response functions and the estimated burden of adverse health effects attributed to exposure concentrations. Risk Anal. 36:1770–1779.

Cox LA Jr, Popken DA. 2015. Has reducing fine particulate matter and ozone caused reduced mortality rates in the United States? Ann Epidemiol. 25:162–173.

Cromar KR, Gladson LA, Perlmutt LD, Ghazipura M, Ewart GW. 2016. American Thoracic Society and Marron Institute Report. Estimated excess morbidity and mortality caused by air pollution above American Thoracic Society-Recommended Standards, 2011–2013. Ann Am Thorac Soc. 13:1195–1201.

Dash D, Voortman M, de Jongh M. 2013. Sequences of mechanisms for causal reasoning in artificial intelligence. Paper presented at: Proceeding IJCAI '13 Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence; 03–09 Aug 2013. Beijing: AAAI Press; p. 839–845.

Dell LD, Mundt KA, Luippold RS, Nunes AP, Cohen L, Burch MT, Heidenreich MJ, Bachand AM; International Carbon Black Association. 2006. A cohort mortality study of employees in the U.S. carbon black industry. J Occup Environ Med. 48:1219–1229.

Dockery DW, Pope CA 3rd, Xu X, Spengler JD, Ware JH, Fay ME, Ferris BG Jr, Speizer FE. 1993. An association between air pollution and mortality in six US cities. N Engl J Med. 329:1753–1759.

Dockery DW, Rich DQ, Goodman PG, Clancy L, Ohman-Strickland P, George P, Kotlov T; HEI Health Review Committee. 2013. Effect of air pollution control on mortality and hospital admissions in Ireland. Res Rep Health Eff Inst. 176:3–109.

Dominici F, Greenstone M, Sunstein CR. 2014. Science and regulation. Particulate matter matters. Science. 344:257–259.

Faes L, Porta A, Nollo G. 2015. Algorithms for the inference of causality in dynamic processes: Application to cardiovascular and cerebrovascular variability. Conf Proc IEEE Eng Med Biol Soc. 2015:1789–1792.

Fann N, Lamson AD, Anenberg SC, Wesson K, Risley D, Hubbell BJ. 2012. Estimating the national public health burden associated with exposure to ambient PM2.5 and ozone. Risk Anal. 32:81–95.

Fedak KM, Bernal A, Capshaw ZA, Gross S. 2015. Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. Emerg Themes Epidemiol. 12:14.

Frangakis CE, Rubin DB. 2002. Principal stratification in causal inference. Biometrics. 58:21–29.

Franklin M, Zeka A, Schwartz J. 2006. Association between PM2.5 and all-cause and specific-cause mortality in 27 US communities. J Expo Sci Environ Epidemiol. 17:279–287.

Frey HC. 2016 Sep. Dose–response models are conditional on determination of causality. Risk Anal. 36:1751–1754.

Frey L, Fisher D, Tsamardinos I, Aliferis CF, Statnikov A. (2003). Identifying Markov blankets with decision tree induction. Paper presented at: Proceedings of the Third IEEE International Conference on Data Mining; 19-22 Nov 2003, Melbourne; p. 59–66.

Friston K, Moran R, Seth AK. 2013. Analysing connectivity with Granger causality and dynamic causal modelling. Curr Opin Neurobiol. 23:172–178.

Furqan MS, Siyal MY. 2016. Random forest Granger causality for detection of effective brain connectivity using high-dimensional data. J Integr Neurosci. 15:55–66.

Giannadaki D, Lelieveld J, Pozzer A. 2016. Implementing the US air quality standard for PM2.5 worldwide can prevent millions of premature deaths per year. Environ Health. 15:88.

Glaeser EL. 2006. Researcher incentives and empirical methods. NBER Technical Working Paper No. 329.

Goto D, Ueda K, Ng CFS, Takami A, Ariga T, Matsuhashi K, Nakajima T. 2016. Estimation of excess mortality due to long-term exposure to PM2.5 in Japan using a high-resolution model for present and future scenarios. Atmos Environ. 140:320–332.

Granger CWJ. 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica. 37:424–438.

Green LC, Crouch EA, Ames MR, Lash TL. 2002. What's wrong with the National Ambient Air Quality Standard (NAAQS) for fine particulate matter (PM(2.5))? Regul Toxicol Pharmacol. 35:327–337.

Greenland S. 2005. Multiple-bias modelling for analysis of observational data. J R Statist Soc A. 168:267–306.

Greven S, Dominici F, Zeger S. 2011. An approach to the estimation of chronic air pollution effects using spatio-temporal information. J Amer Stat Assoc. 106:396–406.

Höfler M. 2005. The Bradford Hill considerations on causality: a counterfactual perspective. Emerg Themes Epidemiol. 2:11.

Höfler M, Lieb R, Wittchen HU. 2007. Estimating causal effects from observational data with a model for multiple bias. Int J Methods Psychiatr Res. 16:77–87.

Halliday DM, Senik MH, Stevenson CW, Mason R. 2016. Non-parametric directionality analysis – extension for removal of a single common predictor and application to time series. J Neurosci Methods. 268:87–97.

Hart J, Garshick E, Dockery D, Smith T, Ryan L, Laden F. 2011. Long-term ambient multipollutant exposures and mortality. Am J Respir Crit Care Med. 183:73–78.

Hernan M, Vanderweele T. 2011. Compound treatments and transportability of causal inference. Epidemiology. 22:368–377.

Hernandez B, Raftery AE, Pennington SR, Parnell AC. 2015 Bayesian additive regression trees using Bayesian model averaging [Internet]. [cited 2017 Mar 30]. Available from: https://arxiv.org/pdf/1507.00181.pdf

Hill AB. 1965. The environment and disease: association or causation? Proc R Soc Med. 58:295–300.

Hill J. 2016. Atlantic Causal Inference Conference Competition: Is your SATT where it's at? [Internet]. [cited 2017 Mar 30]. Available from: http://jenniferhill7.wixsite.com/acic-2016/competition

Imai K, Keele L, Tingley D, Yamamoto T. 2011. Unpacking the black box of causality: learning about causal mechanisms from experimental and observational studies. Am Political Sci Rev. 105:765–789.

Keele L, Tingley D, Yamamoto T. 2015. identifying mechanisms behind policy interventions via causal mediation analysis. J Policy Anal Manage. 34:937–963.

Kenny DA. 1979. Correlation and causality. New York: John Wiley & Sons.

Kim J, Yoon K, Choi JC, Kim H, Song JK. 2016. The association between wind-related variables and stroke symptom onset: a case-crossover study on Jeju Island. Environ Res. 150:97–105.

Kleinberg S, Hripcsak G. 2011. A review of causal inference for biomedical informatics. J Biomed Inform. 44:1102–1112.

Koller D, Friedman N. 2009. Probabilistic Graphical Models. Cambridge (MA): MIT Press.

Krstić G. 2013. A reanalysis of fine particulate matter air pollution versus life expectancy in the United States. J Air Waste Manage Assoc. 63:133–135.

Krstić G, Krstić NS, Zambrano-Bigiarini M. 2016. The br2 – weighting Method for Estimating the Effects of Air Pollution on Population Health. J Mod Appl Stat Methods. 15:723–736.

Laden F, Schwartz J, Speizer FE, Dockery DW. 2006. Reduction in fine particulate air pollution and mortality: extended follow-up of the Harvard Six Cities study. Am J Respir Crit Care Med. 173:667–672.

Lagani V, Triantafillou S, Ball G, Tegnér J, Tsamardinos I. 2016. Probabilistic computational causal discovery for systems biology. In: Geris L, Gomez-Cabrero D, editors. Uncertainty in biology: a computational modeling approach. Studies in Mechanobiology, Tissue Engineering and Biomaterials. Vol. 17. Switzerland: Springer International Publishing; p. 33–73.

Lee S, Honavar V. 2013. m-Transportability: Transportability of a causal effect from multiple environments. Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, Bellevue, Washington, USA.

Lepeule J, Laden F, Dockery D, Schwartz J. 2012. Chronic exposure to fine particles and mortality: an extended follow-up of the Harvard Six Cities study from 1974 to 2009. Environ Health Perspect. 120:965–970.

Lewis-Beck MS, Bryman A, Liao TF (editors). 2003. Encyclopedia of social science research methods. Thousand Oaks (CA): Sage Publications.

Lin H, Liu T, Fang F, Xiao J, Zeng W, Li X, Guo L, Tian L, Schootman M, Stamatakis KA, Qian Z, Ma W. 2016. Mortality benefits of vigorous air quality improvement interventions during the periods of APEC Blue and Parade Blue in Beijing, China. Environ Pollut. pii:S0269-7491(16)31292-1

Lo WC, Shie RH, Chan CC, Lin HH. 2016. Burden of disease attributable to ambient fine particulate matter exposure in Taiwan. J Formos Med Assoc. 116:32–40.

Lopiano KK, Smith RL, Young SS. 2015. Air quality and acute deaths in California, 2000–2012 [Internet]. [cited 2017 Mar 30]. Available from: https://arxiv.org/abs/1502.03062

Maldonado G. 2013. Toward a clearer understanding of causal concepts in epidemiology. Ann Epidemiol. 23:743–749.

Marra G, Radice R, Missiroli S. 2014. Testing the hypothesis of absence of unobserved confounding in semiparametric bivariate probit models. Comput Stat. 29:715.

McClellan RO. 2016. Providing context for ambient particulate matter and estimates of attributable mortality. Risk Anal. 36:1755–1765.

Mohammad Y, Nishida T. 2011. Discovering causal change relationships between processes in complex system. Paper presented at: Institute of Electrical & Electronics Engineers (IEEE), Kyoto, Japan.

Moolgavkar SH. 2016. Fine particulate matter pollution and mortality. Risk Anal. 36:1766–1769.

Morabito M, Crisci A, Messeri A, Capecchi V, Modesti PA, Gensini GF, Orlandini S. 2014. Environmental temperature and thermal indices: what is the most effective predictor of heat-related mortality in different geographical contexts? Scientific World J. 2014:961750.

North DW. 2016. Introduction to special issue on air pollution health risks. Risk Anal. 36:1688–1692.

O'Malley AJ. 2012. Instrumental variable specifications and assumptions for longitudinal analysis of mental health cost offsets. Health Serv Outcomes Res Methodol. 12:254–272.

Obenchain RL, Young SS. 2017. Local control strategy: simple analyses of air pollution data can reveal heterogeneity in longevity outcomes. Risk Anal. Available from: https://doi.org/10.1111/risa.12749

Pearl J. 1993. Comment: graphical models, causality, and intervention. Stat Sci. 8:266–269.

Pearl J. 2000. Causality: Models, reasoning, and inference. 1st edn. New York: Cambridge University Press.

Pearl J. 2009a. Causal inference in statistics: an overview. Stat Surv. 3:96–146.

Pearl J. 2009b. Causality: models, reasoning and inference. 2nd edn. New York: Cambridge University Press.

Pearl J. 2010. An introduction to causal inference. Int J Biostat. 6:7.

Pearl J. 2014. Reply to commentary by Imai, Keele, Tingley, and Yamamoto concerning causal mediation analysis. Psychol Methods. 19:488–492.

Petersen ML, van der Laan MJ. 2014. Causal models and learning from data: integrating causal modeling and statistical estimation. Epidemiology. 25:418–426.

Pope CA III, Cropper M, Coggins J, Cohen A. 2015. Health benefits of air pollution abatement policy: role of the shape of the concentration–response function. J Air Waste Manag Assoc. 65:516–522.

Pope CA III, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, Thurston GD. 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA. 287:1132–1141.

Rich DQ. 2017. Accountability studies of air pollution and health effects: lessons learned and recommendations for future natural experiment opportunities. Environ Int. 100:62–78.

Robins JM, Blevins D, Ritter G, Wulfsohn M. 1992. G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients. Epidemiology. 3:319–336.

Rosenbaum PR, Rubin DB. 1983. The central role of the propensity score in observational studies for causal effects. Biometrika. 70:41–55.

Rottman BM, Hastie R. 2014. Reasoning about causal relationships: inferences on causal networks. Psychol Bull. 140:109–139.

Rubin DB. 1978. Bayesian inference for causal effects: the role of randomization. Ann Stat. 6:34–58.

Rubin DB. 2004. Direct and indirect causal effects via potential outcomes (with discussion). Scand J Stat. 31:161–170.

Rubin DB. Oct 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 66:688–701.

Schiatti L, Nollo G, Rossato G, Faes L. 2015. Extended Granger causality: a new tool to identify the structure of physiological networks. Physiol Meas. 36:827–843.

Schreiber T. 2000. Measuring Information Transfer. Physical Rev Lett. 85:461–464.

Schwartz J. 1994. Air pollution and daily mortality: a review and meta analysis. Environ Res. 64:36–52.

Schwartz J. 2004. The effects of particulate air pollution on daily deaths: a multi-city case crossover analysis. Occup Environ Med. 61:956–961.

Schwartz J, Austin E, Bind MA, Zanobetti A, Koutrakis P. 2015. Estimating causal associations of fine particles with daily deaths in Boston. Am J Epidemiol. 182:644–650.

Schwartz J, Bind MA, Koutrakis P. 2017. Estimating causal effects of local air pollution on daily deaths: effect of low levels. Environ Health Perspect. 125:23–29.

Schwartz J, Dockery DW, Neas LM. 1996. Is daily mortality associated specifically with fine particles? J Air Waste Manag Assoc. 46:927–939.

Schwartz J, Laden F, Zanobetti A. 2002. The concentration–response relation between PM(2.5) and daily deaths. Environ Health Perspect. 110:1025–1029.

Schwartz S, Gatto NM, Campbell UB. 2011. Transportability and causal generalization. Epidemiology. 22:746.

Shadish W, Cook T, Campbell D. 2002. Experimental and quasi-experimental designs for generalized causal inference Boston: Houghton Mifflin.

Shpitser I, Pearl J. 2008. Complete identification methods for the causal hierarchy. J Machine Learn Res. 9:1941–1979.

Simon HA. Causal ordering and identifiability. In: Hood WC, Koopmans TC, editors. Studies in econometric method. Cowles commission for research in economics monograph no. 14. New York: John Wiley & Sons, Inc; p. 49–74.

Simon HA. 1954. Spurious correlation: a causal interpretation. J Am Stat Assoc. 49:467–479.

Slack MK, Draugalis JR. 2001. Establishing the internal and external validity of experimental studies. Am J Health Syst Pharm. 58:2173–2181.

Smith AE. 2016. Inconsistencies in risk analyses for ambient air pollutant regulations. Risk Anal. 36:1737–1744.

Spirtes P. 2010. Introduction to causal inference. J Machine Learn Res. 11:1643–1662.

Spirtes P, Meek C, Richardson T. 1995. Causal inference in the presence of latent variables and selection bias. Paper presented at: UAI'95 Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence; 18–20 Aug 1995, Montréal, Canada. San Francisco (CA): Morgan Kaufmann Publishers Inc; p. 499–506.

Spirtes P, Zhang K. 2016. Causal discovery and inference: concepts and recent methodological advances. Appl Inform (Berl). 3:3.

Sun J, Taylor D, Boll EM. 2015. Causal network inference by optimal causation entropy. SIAM J Appl Dynam Syst. 14:73–106.

Textor J. 2015. Drawing and analyzing causal DAGS with DAGitty [Internet]. [cited 2017 Mar 30]. Available from: http://www.dagitty.net/manual-2.x.pdf

Textor J, Hardt J, Knüppel S. 2011. DAGitty: a graphical tool for analyzing causal diagrams. Epidemiology. 22:745.

United States Environmental Protection Agency (US EPA). 2015. BenMAP environmental benefits mapping and analysis program – community edition. User's manual appendices [Internet]. [cited 2017 Mar 30]. Available from: www.epa.gov/sites/production/files/2015-04/documents/benmap-ce_user_manual_appendices_march_2015.pdf

Urban A, Kyselý J. 2014. Comparison of UTCI with other thermal indices in the assessment of heat and cold effects on cardiovascular mortality in the Czech Republic. Int J Environ Res Public Health. 11:952–967.

Valberg PA. 2003. Possible noncausal bases for correlations between low concentrations of ambient particulate matter and daily mortality. Nonlinearity Biol Toxicol Med. 1:521–530.

Vanderweele TJ. 2011. Principal stratification – uses and limitations. Int J Biostat. 7:28. Available from: https://doi.org/10.2202/1557-4679.1329

Wang C, Tu Y, Yu Z, Lu R. 2015. PM2.5 and cardiovascular diseases in the elderly: an overview. Int J Environ Res Public Health. 12:8187–8197.

Wang Y, Kloog I, Coull BA, Kosheleva A, Zanobetti A, Schwartz JD. 2016. Estimating causal effects of long-term PM2.5 exposure on mortality in New Jersey. Environ Health Perspect. 124:1182–1188.

Wibral M, Pampu N, Priesemann V, Siebenhühner F, Seiwert H, Lindner M, Lizier JT, Vicente R. 2013. Measuring information-transfer delays. PLoS One. 8:e55809.

Wiener N. 1956. The theory of prediction. In: Beckenbach EF, editor. Modern mathematics for engineers. New York: McGraw-Hill.

Woodward J. 2013. Causation and manipulability. In: Zalta EN, editor. The stanford encyclopedia of philosophy. Winter 2016 edn. Available from: https://plato.stanford.edu/archives/win2016/entries/causation-mani/

Wright S. 1921. Correlation and causation. J Agric Res. 20:557–585.

Wu MH, Frye RE, Zouridakis G. 2011. A comparison of multivariate causality based measures of effective connectivity. Comput Biol Med. 41:1132–1141.

Yang H, Testa JR, Carbone M. 2008. Mesothelioma epidemiology, carcinogenesis and pathogenesis. Curr Treat Options Oncol. 9:147–157.

Young SS, Xia JQ. 2013. Assessing geographic heterogeneity and variable importance in an air pollution data set. Stat Anal Data Mining. 6:375–386.

Yule GU. 1926. Why do we sometimes get nonsense-correlations between time-series? A study in sampling and the nature of time-series. J R Stat Soc. 89:1–63.

Zhang JL, Rubin DB. 2003. Estimation of causal effects via principal stratification when some outcomes are truncated by "death". J Educ Behav Stat. 28:353–368.

Zigler CM, Dominici F. 2014. Point: clarifying policy evidence with potential-outcomes thinking–beyond exposure–response estimation in air pollution epidemiology. Am J Epidemiol. 180:1133–1140.

Zigler CM, Dominici F, Wang Y. 2012. Estimating causal effects of air quality regulations using principal stratification for spatially correlated multivariate intermediate outcomes. Biostatistics. 13:289–302.

Zigler CM, Kim C, Choirat C, Hansen JB, Wang Y, Hund L, Samet J, King G, Dominici F; HEI Health Review Committee. 2016. Causal inference methods for estimating long-term health effects of air quality regulations. Res Rep Health Eff Inst. 187:5–49.

Zu K, Tao G, Long C, Goodman J, Valberg P. 2015. Long-range fine particulate matter from the 2002 Quebec forest fires and daily mortality in Greater Boston and New York City. Air Qual Atmos Health. 9:213–221.