Global Epidemiology 2 (2020) 100033

Contents lists available at ScienceDirect

Global Epidemiology

journal homepage: https://www.journals.elsevier.com/global-epidemiology

Research paper

Should air pollution health effects assumptions be tested? Fine particulate matter and COVID-19 mortality as an example

Louis Anthony Cox Jr *, Douglas A. Popken

Cox Associates LLC, 503 N. Franklin Street, Denver, CO 80218, United States of America

ARTICLE INFO

Article history:
Received 3 June 2020
Received in revised form 10 July 2020
Accepted 8 August 2020
Available online 2 September 2020

Keywords:
Air pollution
COVID-19 mortality risk
PM2.5
Health effects
Causation
Scientific method
Model specification error
Regression
Bayesian networks
CART trees
Machine learning
Random forest

ABSTRACT

In the first half of 2020, much excitement in news media and some peer reviewed scientific articles was generated by the discovery that fine particulate matter (PM2.5) concentrations and COVID-19 mortality rates are statistically significantly positively associated in some regression models. This article points out that they are non-significantly negatively associated in other regression models, once omitted confounders (such as latitude and longitude) are included. More importantly, positive regression coefficients can and do arise when (generalized) linear regression models are applied to data with strong nonlinearities, including data on PM2.5, population density, and COVID-19 mortality rates, due to model specification errors. In general, statistical modeling accompanied by judgments about causal interpretations of statistical associations and regression coefficients – the current weight-of-evidence (WoE) approach favored in much current regulatory risk analysis for air pollutants – is not a valid basis for determining whether or to what extent risk of harm to human health would be reduced by reducing exposure. The traditional scientific method based on testing predictive generalizations against data remains a more reliable paradigm for risk analysis and risk management.

Introduction: Scientific Method and Weight-of-Evidence Consensus Judgments as Paradigms for Regulatory Risk Analysis

A recent commentary opined that the science of public health risk assessment for air pollution-associated health effects can and should be advanced by applying traditional scientific methods and principles of sound science for observational data to epidemiological data on air pollution and public health, in order to better understand how much changing exposures changes probabilities of adverse health outcomes [1]. The advocated principles of sound science included the following:

1. *Use conceptually and operationally clear definitions* of exposure and response variables and of the causal effects of interest. For example, do claimed causal relationships refer to direct or to total effects of changes in ambient air pollution levels on changing health risks? When numerical estimates are given for the changes in health risks that are projected to be caused by a change in air pollution, what is assumed about the levels of other causally relevant variables, such as co-pollutants or co-morbidity or income?
2. *Show explicit, independently verifiable derivations* of causal conclusions from stated assumptions and data.
3. *Provide careful qualification of causal interpretations and conclusions* to correctly and transparently characterize remaining uncertainties and ambiguities.
4. Most importantly, *show results from empirical tests of causal conclusions (predictive generalizations)* implied by causal theories or models against observational data.

In our view, testing falsifiable theory-based predictions against data not used in deriving the theory is the sine qua non of traditional sound science as applied in disciplines from astronomy to zymology (e.g., [2]), and we know of no clear methodological reason not to apply it in air pollution health effects research. One purpose of this paper is to discuss and illustrate how nonparametric and graphical (Bayesian network) methods can help to implement this approach in practice, taking as an illustrative example the question of whether a data set provides evidence that past levels of exposure to fine particulate matter (PM2.5) air pollution increase risks of COVID-19-associated mortality.

This approach was contrasted with a popular alternative framework, widely favored in regulatory risk assessment and policy making over the past decade, in which scientists use their best judgments – often said to reflect the (never precisely defined) "weight of evidence" (WoE) from all sources that they consider – to draw causal conclusions

* Corresponding author.
E-mail address: tcoxdenver@aol.com (L.A. Cox).

https://doi.org/10.1016/j.gloepi.2020.100033
2590-1133/© 2020 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

and make policy recommendations. In this view, regulatory risk assessments and review processes should focus on building consensus about the need to regulate by agreeing on expert judgments about the most appropriate causal interpretations and policy implications of statistical "relationships" and "links" between air pollution and health. The epidemiological "links" in question usually refer to positive exposure concentration-response (C-R) regression coefficients in (selected) regression equations. There is no formal requirement in the WoE framework that further technical details or explanations of these coefficients be specified, such as whether or to what extent they reflect residual confounding, model specification errors, measurement errors, or other non-causal factors, before considering them as evidence for a causal relationship in "causal determination" judgments; thus, for example, finding that multiple high-quality observational studies show an association, even though copollutant exposures are difficult to address, exemplifies the support needed for a judgment of "likely to be a causal relationship" in a WoE framework [3]. The WoE approach does not require that causal judgments have more precise conceptual or operational meanings (e.g., distinguishing between necessary, sufficient, or contributing causes; or between direct and indirect effects; or providing an explicit philosophical or logical basis for defining causal effect); or make unambiguous predictions (e.g., about whether or by how much reducing air pollution levels would reduce health risks, given levels of other causally relevant variables); or that such predictions be tested against data before the conclusions are accepted and used to make policy recommendations. To the contrary, advocates of the WoE approach have objected that upholding these principles for air pollution health effects research would "place a nearly unattainable burden of proof" on a community not accustomed to having to provide such empirical proof for its convictions [4].

A second purpose of this paper is to argue that the WoE framework's use of expert judgments about the causal interpretation of significant positive regression coefficients is unnecessary and unsound. It is unnecessary because less restrictive nonparametric techniques allow identification of mutual information (i.e., statistical dependence) between variables without imposing the assumptions of parametric regression models, which invite the risk of model specification errors. It is unsound because significant positive regression coefficient arise in numerous non-causal ways, including model specification error, measurement error, residual confounding, and non-random subject selection; and the regression models and coefficients themselves provide no basis for judging why one is significantly positive, or whether it would remain so if model specification errors and other errors were removed. Human judgment cannot overcome this statistical limitation: if the information that is logically necessary to ascertain causality is missing, judgment alone cannot provide it. However, if direct causes provide unique information about their effects, nonparametric tests for mutual information (or, conversely, for conditional independence) between random variables provide data-driven tests for potential evidence of causality – no mutual information, no evidence of potential causality between exposure and response – without the necessity of judging why regression coefficients might (or might not) be positive (e.g., [5,6]). Such nonparametric information-based tests are robust to many forms of distortion and measurement error that could invalidate more restrictive parametric modeling assumptions (ibid). The second half of this paper seeks to illustrate the use of nonparametric and information-based (conditional independence test) methods for testing for potential evidence of causality in the PM2.5-COVID-19 example.

North [7] framed these two approaches as a clash of paradigms for how best to use data to reach causal conclusions. We view the first paradigm as the traditional scientific method used in most areas of applied science, and consider its demands for tests of assumptions and conclusions against data to be an essential part of this paradigm. Recent methods of causal analysis have led to increased ability to meet these demands using observational data [8,9]. Specifically, we propose and illustrate the following data analysis steps for implementing the four proposed principles of scientific approach and applying them to assess whether a data set (or more than one) provides evidence that exposures (e.g., to PM2.5) increase risk of an adverse response (e.g., COVID-19 mortality):

- *Use conceptually and operationally clear definitions* of exposure and response variables and of the causal effects of interest between them. We accept without change the definitions of PM2.5 exposure concentrations and COVID-19 mortality in the data sources used for our example. We propose that *to be of interest, a causal effect of exposure on response must satisfy the condition that the response is not conditionally independent of exposure*, given the values of other covariates. This is a deliberately expansive constraint, intended to reflect a necessary rather than a sufficient condition; it allows for predictive causation (changes in exposure help to predict subsequent changes in response); manipulative causation (changes in exposure change response probabilities); necessary causation (response probability does not change unless exposure changes); sufficient causation (response probability changes if exposure changes); and various types of contributing causation (response probability changes based on the values of exposure and other variables) [8]. What it does *not* allow for is that an association between PM2.5 and COVID-19 mortality can be judged to be "causal," or to provide evidence of a causal or likely causal relationship, if COVID-19 is conditionally independent of PM2.5 given the values of other variables (e.g., winter temperatures, which might affect both PM2.5 and COVID-19 mortality rates).
- *Show explicit, independently verifiable derivations* of causal conclusions from stated assumptions and data. To implement this requirement, we use standard software packages (e.g., nonparametric classification and regression tree (CART) software and Bayesian network (BN)-learning software) to perform conditional independence tests (e.g., [8,10,11]). In general, the causal conclusions derived by applying conditional independence tests to observational data are as follows: If the null hypothesis of conditional independence between exposure and response (here, PM2.5 and COVID-19 mortality risk) is not rejected (i.e., if there is no arrow between them in a BN, and if PM2.5 is not identified as a significant predictor of COVID-19 mortality risk after conditioning on other variables such as winter temperatures in a CART tree), then these tests provide no evidence that exposure is a cause of response, in the rather permissive sense just discussed. (Of course, pooling data across many individually underpowered studies might allow a more powerful test. Absence of arrows only indicates that no effect was detected and not necessarily that no effect exists; an effect that is too small to be detected cannot be ruled out. However, for large data sets (e.g., [10]), the plausible size of undetected effects is limited, and simulation can be used to put plausible upper bounds on the sizes of hypothesized unobserved effects [12] If the null hypothesis *is* rejected, then the data do provide evidence that that exposure might be a cause of the response (here, that PM2.5 might be a cause of COVID-19 mortality): the proposed necessary condition is satisfied.
- *Provide careful qualification of causal interpretations and conclusions* to correctly and transparently characterize remaining uncertainties and ambiguities. This is done by noting the conditional independence tests are used to test whether a proposed necessary condition for any causal relationship of interest is satisfied. It does neither more nor less. The remaining uncertainties if conditional independence between PM2.5 and COVID-19 mortality risk is *not* rejected are about the sizes of effects that might still exist without having been detected (i.e., without having led to rejection of the null hypothesis of conditional independence). This can be illuminated by studying the smallest effect sizes that are reliably detected. The remaining uncertainties if conditional independence between PM2.5 and COVID-19 mortality risk *is* rejected are about false-positive rates and about why the two variables are not conditionally independent (e.g., does this reflect predictive causation, manipulative association, omitted confounders, or something else).

- *Show results from empirical tests of causal conclusions (predictive generalizations)* implied by causal theories or models against observational data. In this paper, the empirical tests of causal conclusions consist simply of the conditional independence tests for whether COVID-19 mortality risk is conditionally independent of PM2.5, given the values of other variables (e.g., those identified in a CART tree, a random forest ensemble, or via BN learning as predictors of COVID-19 mortality). The predictive generalization that COVID-19 mortality risk should depend on PM2.5 if PM2.5 (if PM2.5 is a direct cause of it) is tested empirically against observational data via conditional independence tests, and the results can be shown explicitly, e.g., as CART trees or BNs with splits or arrows indicating detected dependence relations for which the null hypothesis of conditional independence is rejected based on the observational data.

Thus, we propose that conditional independence tests now widely available in software used throughout much of machine learning, computational statistics, and data science can be used to support the four steps of the scientific approach. By contrast, the portion of the WoE framework on which we focus also has four steps and supporting statistical methods, as discussed and illustrated subsequently; to us, the key part is an expert judgment about whether significant positive C-R regression coefficients should be treated as evidence that reducing exposure would reduce risk, e.g., based on the Bradford Hill considerations (i.e., strength, consistency, temporality, plausibility, etc. of associations) [8].

More generally, our main proposal is that innovations in data science, such as conditional independence tests and supporting software, make such judgments unnecessary, at least to the extent that these tools can be applied reliably to the types and sizes of data sets that are available, which has become increasingly practical with the development and widespread application of relevant machine learning and computational statistics methods and packages in recent years (e.g., [13]; Glymour et al., 2019; [8,10,11]). Instead, the results of testing testable implications of the causal hypothesis that exposure increases risk can be shown as evidence about the extent to which data do or do not support the hypothesis that exposure is a cause of an adverse effect in an exposed population. Many testable implications of the hypothesis that exposures cause adverse health effects, along with principles and algorithms for testing these implications using observational data, have been developed over the past century, and have been shown to work well in practice by various metrics for many simulated and real data sets ([13,14]). Examples include the following [8]:

- *Effects depend on their direct causes.* Conditional independence tests test this by ascertaining whether data allow the corresponding null hypothesis, that an effect is conditionally independent of a hypothesized direct cause to be rejected. If the probability distribution for the effect differs significantly for different values of the hypothesized direct cause, holding other potential direct causes fixed, this provides evidence that the effect depends on the hypothesized direct cause.
- *An effect's direct causes determine its probability distribution.* This leads to Simon-Iwasaki causal ordering algorithms showing which variables must have their values determined first in order to determine the values of other variables [15,16]).
- *The conditional probability distribution for an effect, given the values of all of its direct causes, is the same even in different settings (*i.e.*, even if other variables have different values, or are set to different values).* Formalizing this intuition leads to statistical tests for the property of Invariant Causal Prediction (ICP): that the dependence of an effect on its direct causes (e.g., its conditional probability distribution, given the values of its direct causes) is the same in different environments and under different interventions ([17]).
- *Information flows from causes to their direct effects over time.* Changes in causes help to predict and explain subsequent changes in the probability distribution of their direct effects. Various formalizations of this concept have been developed ([18,19]), recently leading to software implementing nonparametric tests and estimation procedures for information flows between time series variables based on transfer entropy ([20]).

The example in this paper focuses on conditional independence testing, which is relatively well developed and undemanding ([11,14,21]): unlike ICP, it can be applied to a single data set; and unlike transfer entropy, it does not require time series data for both cause and effect. But the larger point is that innovations in data science and computational statistics make it practical to test many proposed implications of causality with observational data ([8,14]), or with a mix of observational and interventional data [22]. Doing so, and displaying the results, advances the application of principles 1–4 above.

North [7] interprets these innovations as a new paradigm for causal modeling. He may well be right, but we also view testing theory-derived predictions against observations using independently verifiable calculations and reproducible procedures as defining elements of scientific method since at least Galileo [23]. By contrast, we view the WoE paradigm's rejection of the need for such empirical tests in favor of the authoritative judgments of selected experts as a retreat from the traditional requirements of sound science. Although the WoE paradigm is sometimes described as an approach to "assessing all the evidence," we focus here on its use of judgment to assess the causal significance of epidemiological evidence consisting of significant positive C-R regression coefficients. This type of "evidence" has played a dominant role in recent claims about adverse health effects attributed to PM2.5, including suggestions that PM2.5 increases COVID-19-related mortality risk [24]. We object to it on the grounds that finding a significant positive C-R regression coefficient typically usually has no implications for the hypothesis of causality, and appealing to judgment cannot fix this limitation of what regression coefficients show (e.g., that conditioning on exposure reduces mean squared prediction error for the response) or make them show something more relevant for causal inference (e.g., whether changing exposure would change the probability distribution of the response) [21]. North traces the divergence of these paradigms to the acceptance into epidemiology and regulatory risk assessment of the influential work of Sir Austen Bradford Hill in the 1960s, which sought a basis for making intuitive judgments about whether epidemiological associations were best explained as being causal, without applying formal causal analysis methods or testing competing explanations. Ironically, those who favor the WoE framework often characterize calls to apply the scientific method as an attack on science, rather than as a challenge to experts to apply science to back up their judgments with science [25]. The frequency and ferocity of ad hominem attacks (e.g., [26]) suggests that North's diagnosis of a clash of paradigms may well be correct.

This article continues the discussion using an important recent real-world example to illustrate how the paradigms differ and why the choice between them matters: interpreting studies associating air pollution and COVID-19.

Avoiding the Burden of Empirical Proof by Using Regression Models and Judgment

Concern that compelling evidence for human health benefits caused by tighter regulation of air pollution might be unattainable from real-world data [4] is well-founded: the benefits assumed and claimed in WoE analyses have proved difficult to find in evidence-based studies that have compared public health risks before and after pollution-reducing interventions or changes [27], even under conditions where they should have been easily seen if they were approximately as large as claimed [12]. For example, for fine particulate matter (PM2.5), a positive correlation between *levels* of PM2.5 and mortality is clear in many studies – both PM2.5 levels and mortality rates are higher in some times and places than in others, inducing strong correlations and regression coefficients between them. This has sufficed to drive causal determinations and recommendations to regulate in a WoE framework that deals in vague "links" and "relationships." But in multiple studies in multiple countries over many years, reducing PM2.5 has not been found to have an unequivocal causal effect – or, in many studies, even a clear association – with *changes* in all-cause mortality risk [10,27]), which would be the hallmark of a genuine causal relationship between them [21]. For example, Burns et al. [27], after reviewing 42 such studies, conclude that "Given the heterogeneity across interventions, outcomes, and methods, it was difficult to derive overall conclusions regarding the effectiveness of interventions in terms of improved air quality or health. Some evidence suggests that interventions are associated with improvements in air quality and human health, with very little evidence suggesting interventions were harmful." Of course, as Burns et al. [27] also emphasize, absence of clear evidence is not clear evidence of absence of an effect, although perhaps it is clear evidence of absence of an effect large enough to detect in the studies reviewed, or of the sizes predicted by regression models when regression coefficients are interpreted causally [12].

The absence of clear evidence that regulations or other interventions that reduce ambient air pollution in recent decades have caused reductions in all-cause mortality has often been met by using expert judgments, regression models of associations, and counterfactual causal interpretations of regression model results to predict that these changes *should* take place in theory (i.e., according to the regression models if they are interpreted causally), whether or not they actually *do* take place. Predictions from computer models stocked with consensus assumptions, rather than empirical validation of predictions against data, are generally treated as sufficient in the WoE paradigm to draw conclusions and policy recommendations to be shared with policymakers and the press. For example, in the United States, the United States Environmental Protection Agency (US EPA) BenMAP-C computer model uses regression models and expert judgments to predict how changes in air pollution would change public health effects, even though the detailed documentation for its health impact functions repeatedly notes that causal information was not included [28]. It is well understood in epidemiology that, technically, correlation is not causality and regression coefficients reflect only whether predictors help to predict the dependent variable in a regression model (e.g., reducing the mean squared error (MSE) or increasing the value of the likelihood function for regression-based predictions), and not whether or how much changing the values of predictors would change the distribution of the dependent variable [21]. Nonetheless, it remains common practice to present estimated or assumed air pollution concentration-health response (C-R) associations and regression coefficients *as if* they were causal relationships with life-and-death implications; users of BenMAP-C often make this assumption [28]. Authoritative expert judgment and consensus bridge the evidentiary gap: a C-R association is treated as if it were causal if appropriate authorities – or the scientists doing the work and reporting the results – agree that they think causality is the best explanation for it [29]. As one recent example among many, Chen et al. [30] estimated that a "reduction in PM2.5 during the [COVID-19] quarantine period avoided a total of 3214 PM2.5-related deaths (95% CI 2340–4087) in China, 73% of which were from cardiovascular diseases" during a 34-day quarantine period. These numbers were calculated by assuming (and therefore predicting) that reducing PM2.5 concentrations causes approximately proportional reductions in daily mortalities, with the constant of proportionality being estimated from statistical associations in past data. Thus, the claim that reducing PM2.5 "avoided a total of 3214 PM2.5-related deaths" is not driven by observations of any actual reduction in death counts compared to what would have been expected in the absence of reduced PM2.5. Rather, it reflects a judgment that previously estimated statistical slope coefficients describing C-R associations should be used to project reduced mortalities. Such judgment-based projections require no observations about actual death counts during the quarantine period.

That comparing model predictions to real-world observations is not necessary for applying the WoE paradigm is also well illustrated by studies in the United States that predict human health benefits from reducing ambient pollution levels. Such predictions can be generated conveniently using the BenMAP-C software, which supplies the judgment-based assumptions and regression models needed to generate positive heath benefits estimates. In defending this use against objections that it treats association as causation, assumptions as data, and hypothetical predictions as facts [28], proponents explained that "The purpose of our report was not to demonstrate causation between exposure to O3 and PM2.5 air pollution and adverse health effects. Our estimates of excess morbidity and mortality are based not simply on observed associations but, rather, on the 'hundreds of epidemiology studies and decades of related scientific research' that clearly establish a relationship between exposure to PM2.5 and O3 and adverse health outcomes. … BenMAP is a well-established research tool that has been used by many investigators to develop estimates of the health benefits that can be achieved by reducing air pollution" [29]. Here, the precedent and popularity of treating an established statistical C-R "relationship" (specifically, positive C-R regression coefficients) between exposure to PM2.5 and O3 and adverse health outcomes as being causal is deemed sufficient justification for continuing the practice. The burden of empirical proof – either showing that the model projections successfully predict real-world experience (e.g., that substantial reductions in PM.5 are followed by corresponding changes in the adverse health effects said to be caused by PM2.5), or else explaining why not and revising the assumptions in the BenMAP model accordingly – is avoided by substituting computer simulations for reality and appealing to expert judgment and tradition to decide whether to accept the simulation results as real for purposes of policy making and risk communication. Failure of the projected benefits to be detected in real data ([10,12]; 26] is of no consequence in a WoE framework that treats expert judgment, precedent, and consensus as the ultimate arbiters for which modeling assumptions and predictions should be accepted. But this also deprives those who rely on the results of the opportunity to learn from reality and to correct errors in modeling assumptions. "Burden of proof" [4] may be too strong a phrase, in that epidemiological papers seldom seek to prove their causal conclusions, but only to present evidence, which is usually less than conclusive. However, the guidance that causal claims should make explicit, empirically testable predictions (such as that effects are not conditionally independent of their direct causes, and other implications discussed previously), and that these predictions should in fact be tested and the results presented before stating causal conclusions, are not onerous to implement, as illustrated next.

Interpreting Regression Models for PM2.5 and COVID-19 Deaths

As COVID-19 mortalities mounted worldwide in the first two quarters of 2020, environmental activists and scientists rushed to shape policy with headlines and scientific articles warning that fine particulate matter air pollution (PM2.5) increases risk of COVID-19-related illness and death. Once again, WoE thinking and unverified model predictions paved the way. For example, Jiang et al. [31] used a Poisson regression model to conclude that PM2.5 and humidity increased the risk of daily COVID-19 incidence in three Chinese cities, while coarse particulate air pollution (PM10) and temperature decreased this risk. Bashir et al. [32] calculated significant ordinal correlations between PM2.5 and other air pollutants (PM10, SO2, NO2, and CO) and COVID-19 cases in California, and concluded that such correlations should encourage regulators to more tightly control pollution sources to prevent harm. Most famously, Wu et al. [24] fit a negative binomial regression model to county-level data in the United States, and interpreted their finding of a significant positive regression coefficient for PM2.5 as implying that "A small increase in long-term exposure to PM2.5 leads to a large increase in the COVID-19 death rate." This interpretation attracted national headlines and widespread political concern (Friedman 2020).

These examples follow a common technical approach with the following steps, which we view as exemplifying WoE thinking as it applies to interpreting evidence from one (or more) regression models:

1. Collect data on estimated air pollution levels, one or more adverse health outcomes of interest (such as COVID-19 mortality), and covariates of interest (e.g., humidity, temperature, population density, etc.)
2. Fit one or more regression model to the data, treating air pollution levels as predictors and adverse health outcomes as dependent variables. Include other variables as covariates at the modeler's discretion.
3. If the regression coefficient for a pollutant as a predictor of an adverse health outcome is significantly positive in the one or more regression models, use judgment to interpret this as evidence that reducing levels of the pollutant would reduce risk of the adverse health outcome.
4. Communicate the results to policy makers and the press using the policy-relevant language of *causation* and *change* – that is, claim that a given reduction in pollution would create a corresponding reduction in adverse health outcomes – rather than in the (technically accurate) language of *association* and *difference*: that a given difference in estimated exposures is associated with a corresponding difference in the conditional expected value of a dependent variable predicted by the selected regression model.

Step 3 is based on a judgment that a positive regression coefficient in a modeler-selected regression model is evidence of a causal relationship: that it implies or suggests that reducing exposure would reduce risk, even if the experiment has not actually been made. In this respect, it incorporates the central principle of the WoE framework: that a well-informed expert scientist can make a useful judgment about whether the association indicated by a statistically significant positive regression coefficient is likely to be causal. We next scrutinize this assumption.

Do Positive Regression Coefficients Provide Evidence of Causation?

As noted by Dominici et al. [33], either significant positive coefficients or significant negative regression coefficients (or no significant regression coefficient at all) for air pollution as a predictor of mortality risk can often be produced from the same data, depending on the modeling choices made; thus "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." In the field of air pollution health effects research, however, investigators continue to rely on regression modeling in step 2 of the above approach. A skilled regression modeler can usually produce a model with a significant positive regression coefficient for exposure in step 2, allowing steps 3 and 4 to proceed. We illustrate next how this can be done, using a data set on PM2.5 and COVID-19 mortality in the United States as an example. The data set, described and provided via a web link in Appendix A, compiles county-level data on historical ambient PM2.5 concentration estimates, COVID-19 mortality rates and case rates (per 100,000 people) through April of 2020, along with numerous other county-level variables.

A key step in regression modeling is to select variables to include in the model. Fig. 1 shows a random forest (nonparametric model ensemble) importance plot for county-level variables as predictors of COVID-19 mortality rates, where the "importance" of each variable is measured by the estimated percentage increase in mean squared prediction error if that variable is dropped as a predictor. The few most important predictors of COVID-19 mortality (*DeathRate100k*) are *PCT_BLACK*, the percentage of a county population that is Black; *PopDensity*, the average density of the population in the county (number of people per square mile) or its logarithm, *PopDensityLog* (the log transform makes little difference to nonparametric methods such as random forest, but can be important for parametric regression models); *Longitude*, time since first case in the county (*FirstCaseDays*), average estimated PM2.5 concentration between 2000 and 2016 (*X2000.2016AveragePM25*), average temperature during the winter

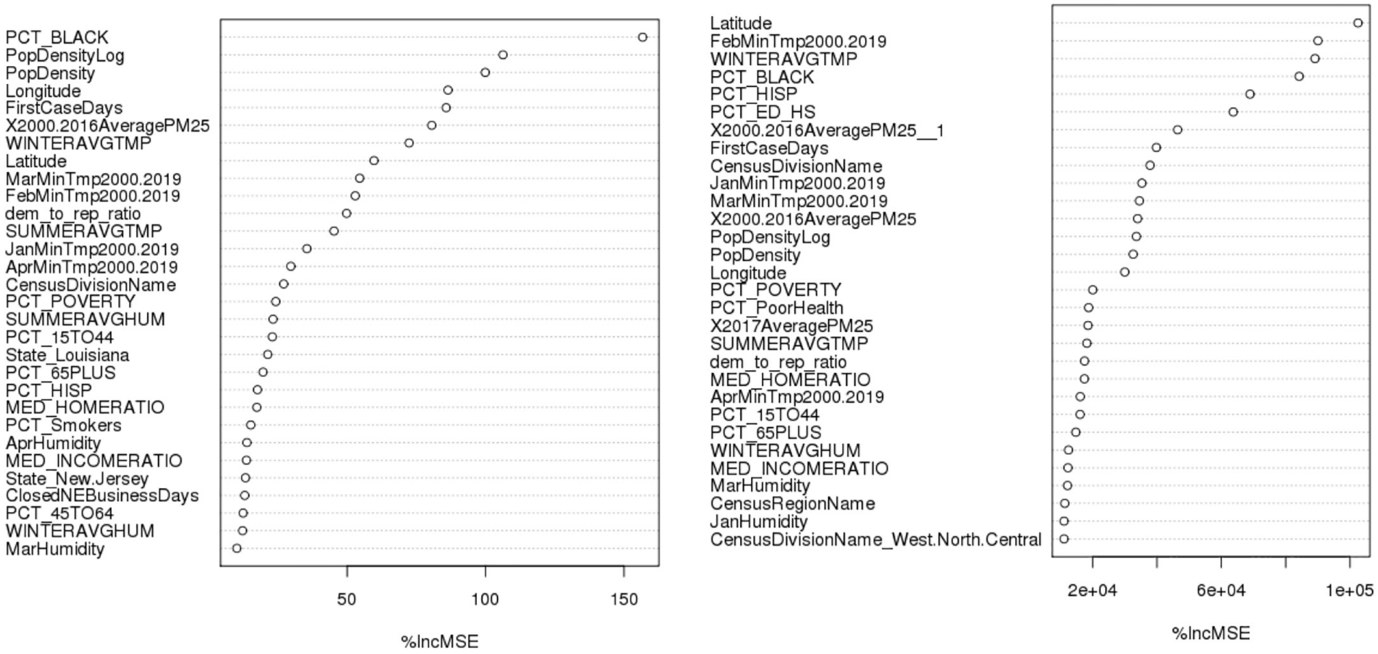


Fig. 1. Importance plots for several variables as predictors of COVID-19 mortality (left) and case rates (right) per 100,000 people. The plots are generated by random forest nonparametric model ensembles that explain about 48% of the variance in mortality rates and 40% of the variability in case rates among counties in the United States as of April 2020. Appendix A provides details and data. "Importance" is measured as the percentage increase in mean squared prediction error ("%IncMSE") if a variable is dropped as a predictor. Variable labels are defined in the text and in Appendix A for the most important variables; see data sources in Appendix A for all variables. "%IncMSE" is the percentage increase in mean squared prediction error from dropping a variable.

months between 2000 and 2016 (*WINTERAVGTMP*), and *Latitude*. For the case rate (COVID-19 cases reported per 100,000 people), the most important predictors also include the average minimum temperature in February over the past two decades (*FebMinTmp2000.2019*), percent Hispanic (*PCT_HISP*), and percent of population with at least high school educations (*PCT_ED_HS*). These ten predictors alone explain about the same percentages of the variances in COVID-19 mortality and case rates across counties (48% and 40%, respectively) as the full set of over 60 variables, of which the most important are shown in Fig. 1.

Of course, predictors need not be statistically independent of each other. To visualize the statistical interdependencies among them, Fig. 2 shows a Bayesian network (BN) fit to the data (using the default hill-climbing (HC) algorithm in the *bnlearn* package in R, see www.bnlearn.com/ and Appendix A), with *Latitude* and *Longitude* constrained to have only outward-pointing arrows and *DeathRate100k* constrained to have only inward-pointing arrows, to facilitate possible intuitive causal interpretations of the arrows leaving or entering these three nodes. (Presumably, latitude and longitude are not caused by anything else, and death does cause any of the other variables.) However, in general the arrows only signify statistical dependencies between variables, and not necessarily causal relationships.

For example, an arrow between PM2.5 and percent Hispanic (*X2000.2016AveragePM25* and *PCT_HISP*) does not suggest that either causes the other: it simply reflects that counties with higher percentages of Hispanic populations tend to also have higher PM2.5 levels. However, if variables depend on their direct causes, then absence of an arrow between two variables corresponds to absence of empirical evidence in the BN that either directly causes the other. COVID-19 mortality in Fig. 2 is shown as depending directly on latitude and longitude (which are presumably surrogates for other biologically effective causes), as well as on time since first case in a county (*FirstCaseDays*), average winter temperature, and ethnic composition (*PCT_BLACK* and *PCT_HISP*). Fig. 3 shows an analogous BN for COVID-19 case rate, which depends directly on latitude and longitude, ethnic composition (*PCT_BLACK* and *PCT_HISP*), time since first case in a county (*FirstCaseDays*), and education (*PCT_ED_HS*).

Bayesian network learning is a relatively new technique for exploring and visualizing direct and indirect dependencies among variables. As an alternative, Fig. 4 shows a classification and regression tree (CART) tree for COVID-19 mortality. The CART algorithm (implemented in the *rpart* package in R) recursively partitions counties into clusters with significantly different COVID-19 mortality rates, based on the results of binary tests ("splits"), such as whether *Longitude* < -75.61 (yes = left branch, no = right branch). For example, the counties with *Longitude* $< - 75.61$, *PCT_BLACK* < 0.2636, and time since first case <44.5 days have an average COVID-19 mortality rate of less than 3 per 100,000 (2.436, although 3 significant digits is spurious precision), compared to a rate over 50 times greater (148.7 per 100,000) for counties further to the East with high population densities and longer times since first cases. Although CART trees are subject to residual confounding due their binary splits of continuous variables and are not very robust, in the sense that fitting them to multiple random samples from the same data set often produces different trees (e.g., with *WINTERAVGTMP* in some and *FebMinTmp2000.2019* in others), they provide a relatively simple, well-established nonparametric technique for exploring significant predictors of a selected dependent variable such as *DeathRate100K*. The predictors identified in Fig. 4 are *Longitude*, *PCT_BLACK*, *WINTERAVGTMP*, *FirstCaseDays*, and *PopDensity*.

Although we regard Figs. 2-4 as only exploratory visualizations, they highlight the importance of confounders such as *Longitude* in understanding associations between PM2.5 and COVID-19 mortality. The usual way to control for measured confounders in regression modeling is to "adjust" for them by including them on the right side of a regression equation. For example, a multiple linear regression model that includes all of the variables identified in Figs. 2 and 4 on which COVID-19 mortality rate might directly depend, and that further hypothesizes a dependence on historical average ambient PM2.5 exposure concentration levels (*X2000.2016AveragePM25*), would posit an equation of the form in eq. (1).

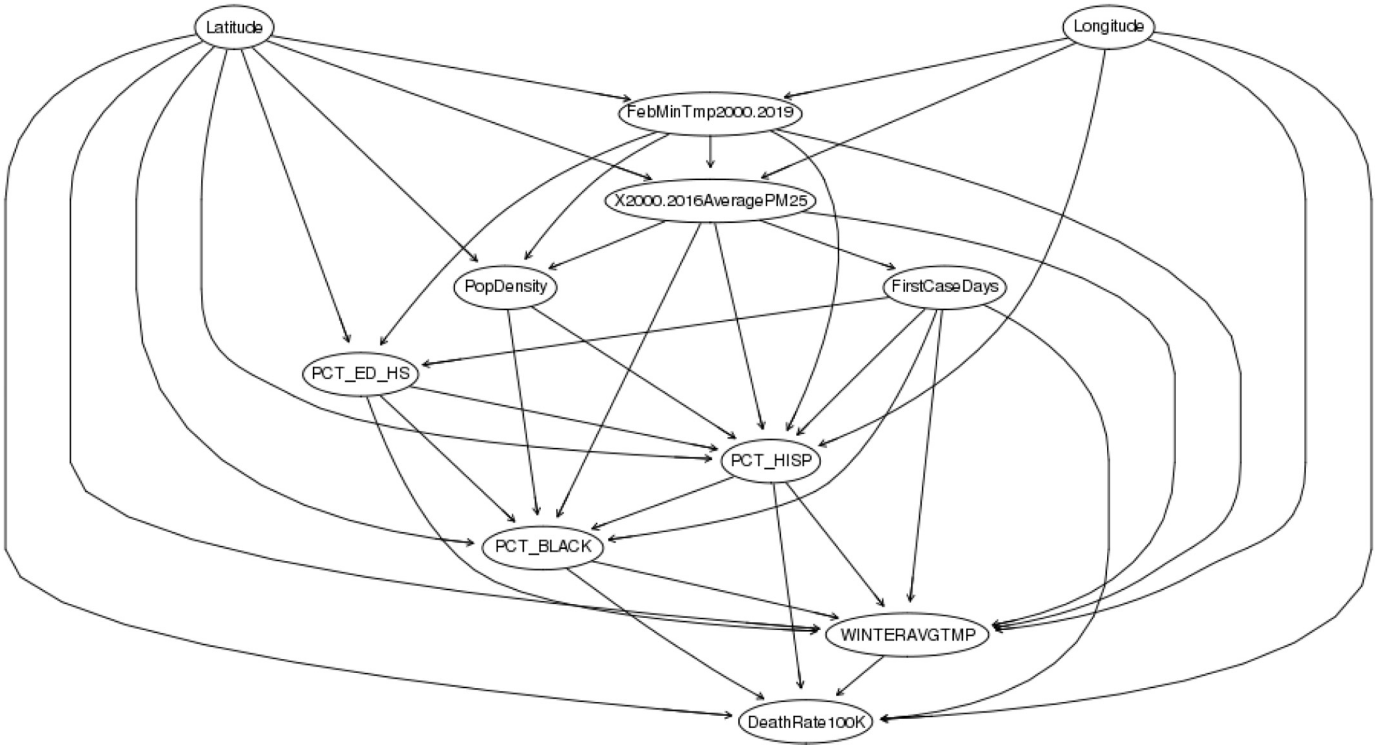


Fig. 2. Bayesian network for COVID-19 mortality (deaths per 100,000 people) showing statistical dependencies among variables. An arrow between two variables indicates that they are informative about each other (i.e., not statistically independent).

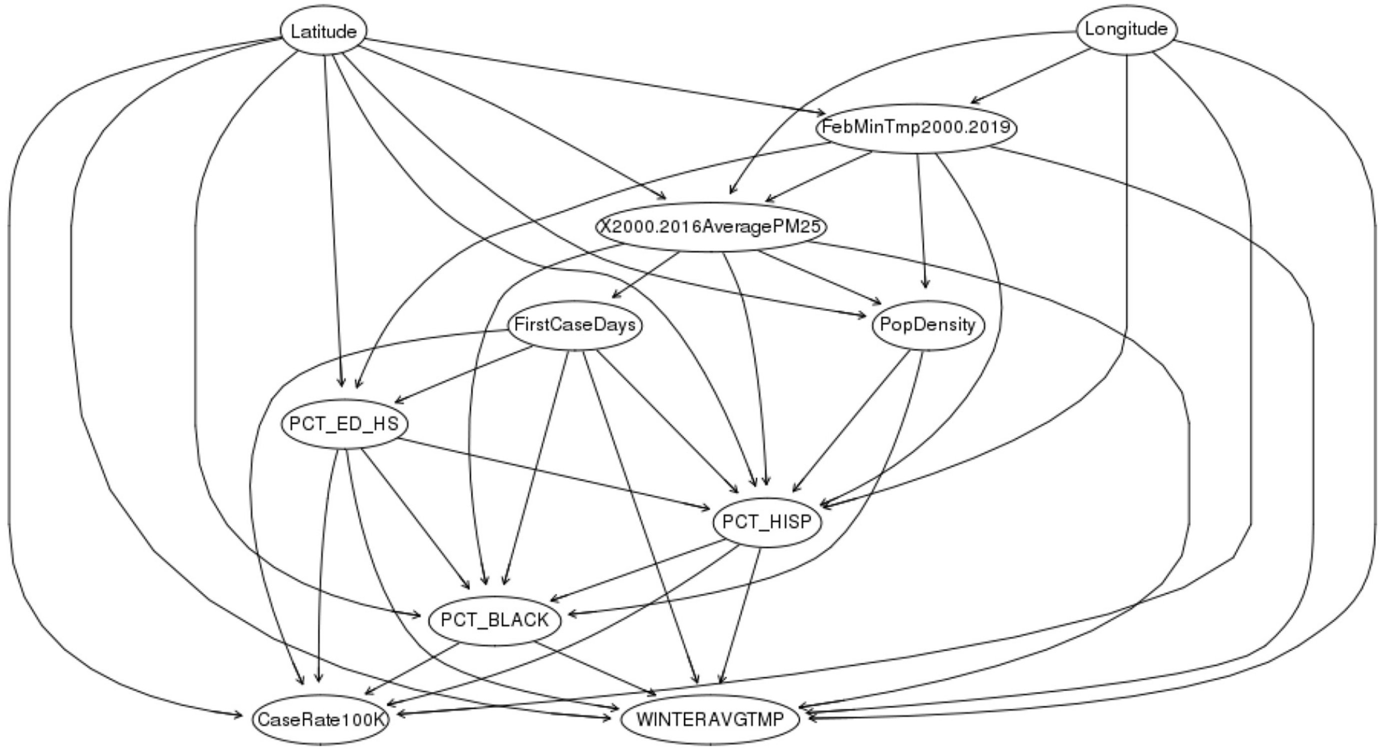


Fig. 3. Bayesian network for COVID-19 case rate (cases reported per 100,000 people).

$$\begin{aligned} E(DeathRate100K) = {} & Intercept + b_1 * X2000.2016AveragePM25 + b_2 \\ & * PCT\_BLACK + b_3 * PCT\_HISP + b_4 * Latitude \\ & + b_5 * Longitude + b_6 * FirstCaseDays + b_7 \\ & * WINTERAVGTMP + b_8 * PopDensity \end{aligned} \qquad (1)$$

For simplicity, eq. (1) follows Wu et al. [24] in assuming that risk depends on a weighted sum of terms on the right side, ignoring interaction terms (e.g., that increasing PM2.5 should not increase death rates if population density = 0); consequences of this modeling choice are discussed later. Fitting eq. (1) to the data set via ordinary least squares (OLS) regression yields Table 1.

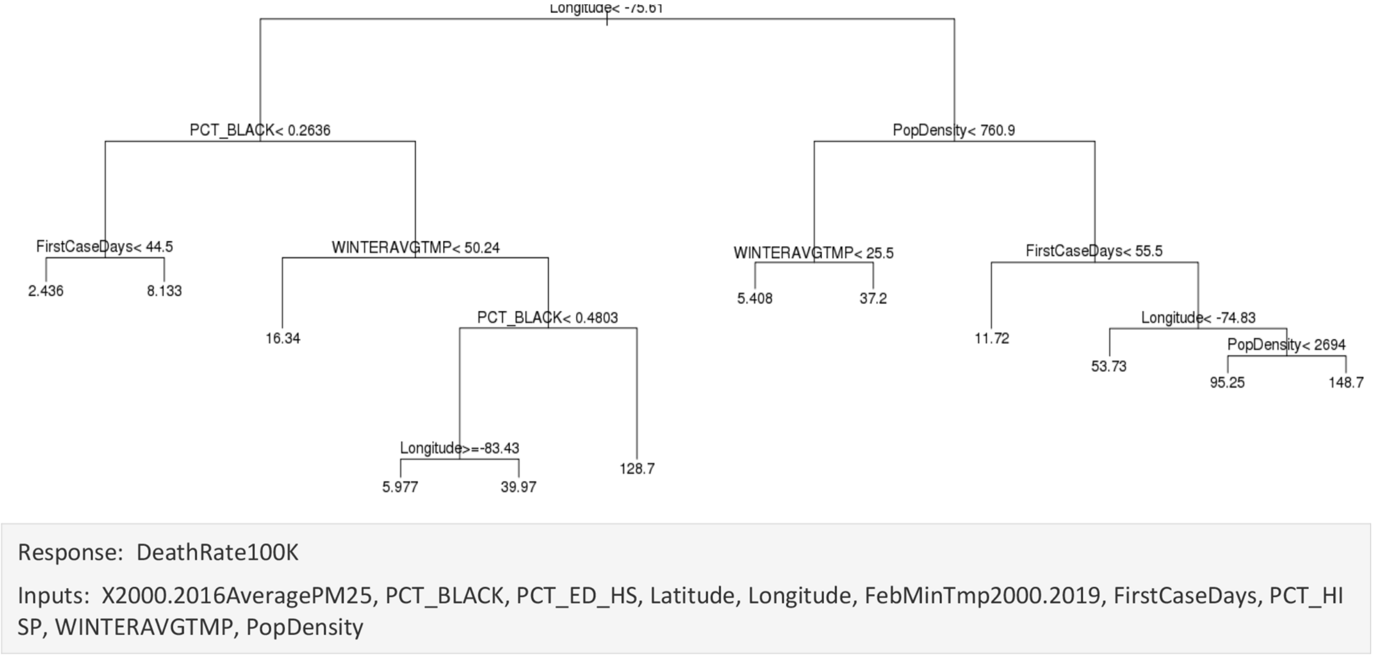
The regression coefficient for past estimated average ambient PM2.5 exposure concentration (denoted by *2000.2016AveragePM25* in Tables 1 and 2) is negative and not significantly different from 0 ($p = 0.87$), consistent with Fig. 2. However, regression modeling allows modelers to select variables to include in the model, which can drive the results that get published [33]. For example, dropping *Longitude* from the regression model yields Table 2. Now the regression coefficient for PM2.5 is positive instead of negative, and it is highly significant ($p = 0.000053$) instead of non-significant. In effect, PM2.5 acts as a partial surrogate for longitude for predicting COVID-19 mortality risk, so that omitting longitude induces PM2.5 to enter with a significant positive coefficient. Fig. 5 suggests why: both PM2.5 and COVID-19 mortality rates tend to be higher in the East than in the West. (COVID-19 cases and death rates in April 2020 were far higher in New York City and adjacent areas of New York, New Jersey, and Connecticut than in most other parts of the United States.) Interpreting the positive regression coefficient for PM2.5 in Table 2 as evidence that an increase in PM2.5 increases PM2.5 mortality risk would be mistaken: it is only evidence that the modelers made choices (such as omitting longitude from the model) that led to a positive regression coefficient.

This example not only illustrates the obvious point that omitting from a regression model predictors such as longitude, on which both PM2.5 and COVID-19 mortality rate depend (Figs. 2, 4 and 5), can induce a significant positive C-R regression coefficient for PM2.5 when COVID-19 mortality rate is regressed against it and other variables; but it also illustrates the more constructive point that nonparametric methods can help to identify variables that must be conditioned on to avoid such spurious C-R coefficients. The BN in Fig. 2 indicates that longitude provides information about both PM2.5 and COVID-19 mortality rates that the other variables do not, implying that it is a potential confounder that must be adjusted for in order to obtain unbiased C-R coefficients [34]).

The linear model (1) was selected for simplicity rather than realism, to illustrate how significant positive C-R regression coefficients can easily arise in (misspecified) parametric regression models even if no corresponding dependencies are found in non-parametric models or in BNs (Figs. 2 and 4). The usual assumptions of the linear regression model (e.g., homoscedasticity, normally distributed additive error terms) would be better justified if the model in eq. (1) were further refined, e.g., by log-transforming the dependent variable (which raises the R^2 value from 0.145 to 0.30). Many additional variables could be included on the right side of the model, as in the work of Wu et al. [24], further improving model fit and increasing R^2. However, such improvements and elaborations would not change the key point that simple parametric regression model forms that do not include all relevant predictors, or that fail to model important interactions and nonlinearities, can thereby create significant positive C-R regression coefficients for exposure that are spurious, in the sense that they do not reflect a dependence of response on exposure, but only reflect model specification errors. The following section examines more carefully how omitting nonlinearities from regression models can create such spurious significant positive C-R coefficients for exposure even if risk does not depend on exposure.

Positive Regression Coefficients Created by Model Specification Error and Other Causes.

More generally, there are many reasons that PM2.5 might have a significant positive regression coefficient that do *not* imply that increasing



Response: DeathRate100K

Inputs: X2000.2016AveragePM25, PCT_BLACK, PCT_ED_HS, Latitude, Longitude, FebMinTmp2000.2019, FirstCaseDays, PCT_HISP, WINTERAVGTMP, PopDensity

Fig. 4. A classification and regression tree (CART) tree for COVID-19 mortality (*DeathRate100K*). The tree was generated by the *rpart* package in R. Response: DeathRate100K. Inputs: X2000.2016AveragePM25, PCT_BLACK, PCT_ED_HS, Latitude, Longitude, FebMinTmp2000.2019, FirstCaseDays, PCT_HISP, WINTERAVGTMP, PopDensity.

PM2.5 would increase risk. As a simple conceptual example, suppose that *PopDensityLog* is a confounder of the PM2.5-Risk association, and that the structural equations describing the causal relationships among these variables are as follows:

$$E(Risk) = PopDensityLog^2 \tag{2}$$

$$PM2.5 = PopDensityLog^2 \tag{3}$$

In other words, risk increases as population density increases (and only as population density increases), and PM2.5 increases as population density increases (and only as population density increases), but increasing PM2.5 alone results in no increase in risk. Both $E(Risk)$ and PM2.5 increase as the square of *PopDensityLog*. (In such structural equations, the values of the dependent variables on the left are causally determined by the values of the variables on the right: if the right-hand variables are exogenously changed, then the left-hand variables will change to make the equality hold.) Then a model that minimizes prediction error is $E(Risk) = PM2.5$; this yields perfect predictions (since, by hypothesis, $PM2.5 = PopDensityLog^2 = E(Risk)$). Equivalently, the following multiple linear regression model (4) with $b_0 = 0$, $b_1 = 0$, and $b_2 = 1$ has zero mean squared error:

$$E(Risk) = b_0 + b_1 * PopDensityLog + b_2 * PM2.5 \tag{4}$$

If model (4) is fit to a large data set, e.g., 1000 observations in which *PopDensityLog* is randomly sampled from a continuous distribution (e.g., with values uniformly distributed between 0 and 1) and corresponding values of $E(Risk)$ and *PM2.5* are calculated using eqs. (2) and (3), respectively, then the ordinary least-squares fit will be $b_0 = 0$, $b_1 = 0$, and $b_2 = 1$. Thus, regression identifies a significant positive coefficient for *PM2.5*, and not for the confounder *PopDensityLog*, because these parameter values minimize prediction error (MSE). But this coefficient has no relevance for determining how or whether changing *PM2.5* would change *Risk*. A claim that such a regression analysis had "controlled for" potential confounding from *PopDensityLog* by including it in regression model (4), and yet still found that *PM2.5* increased risk, would be wrong. A judgment that such an analysis provides evidence that increasing PM2.5 levels increases risk would be mistaken.

As a less hypothetical example, suppose we create a synthetic data set that is identical to the one for Table 2, except for the addition of a

Table 1
Mutiple linear regression model for COVID-19 mortality rate. The columns give standardized regression coefficients (b^*) and their standard errors; unstandardized regression coefficients (b) and their standard errors; and t-test values and significance levels (p-values) for each coefficient.

$N = 3009$	Regression Summary for Dependent Variable: DeathRate100K $R = 0.39594357$ $R^2 = 0.16$ Adjusted $R^2 = 0.16$					
	b*	Std.Err.	b	Std.Err.	t(3000)	p-value
Intercept			−15.81	9.28	−1.7	0.0885
2000-2016AveragePM25	0.00	0.028	−0.04	0.23	−0.2	0.8718
PCT_HISP	0.09	0.021	13.17	3.19	4.1	0.0000
PCT_BLACK	0.28	0.021	40.75	2.99	13.6	0.0000
PopDensity	0.04	0.018	0.00	0.00	2.2	0.0261
WINTERAVGTMP	0.09	0.034	0.17	0.06	2.7	0.0063
FirstCaseDays	0.17	0.018	0.21	0.02	9.6	0.0000
Latitude	0.18	0.037	0.77	0.16	4.8	0.0000
Longitude	0.15	0.025	0.27	0.05	5.9	0.0000

Table 2
Mutiple linear regression model for COVID-19 mortality rate with Longitude omitted.

$N = 3009$	Regression Summary for Dependent Variable: DeathRate100K $R = 0.38341345$ $R^2 = 0.147$ Adjusted $R^2 = 0.145$					
	b*	Std.Err.	b	Std.Err.	t(3001)	p-value
Intercept			−38.08	8.52	−4.5	0.0000
2000-2016AveragePM25	0.09	0.02	0.77	0.19	4.0	0.0001
PCT_HISP	0.06	0.02	9.05	3.13	2.9	0.0038
PCT_BLACK	0.29	0.02	41.08	3.00	13.7	0.0000
PopDensity	0.04	0.02	0.00	0.00	2.5	0.0109
WINTERAVGTMP	0.04	0.03	0.08	0.06	1.3	0.1876
FirstCaseDays	0.17	0.02	0.21	0.02	9.6	0.0000
Latitude	0.15	0.04	0.63	0.16	4.0	0.0001

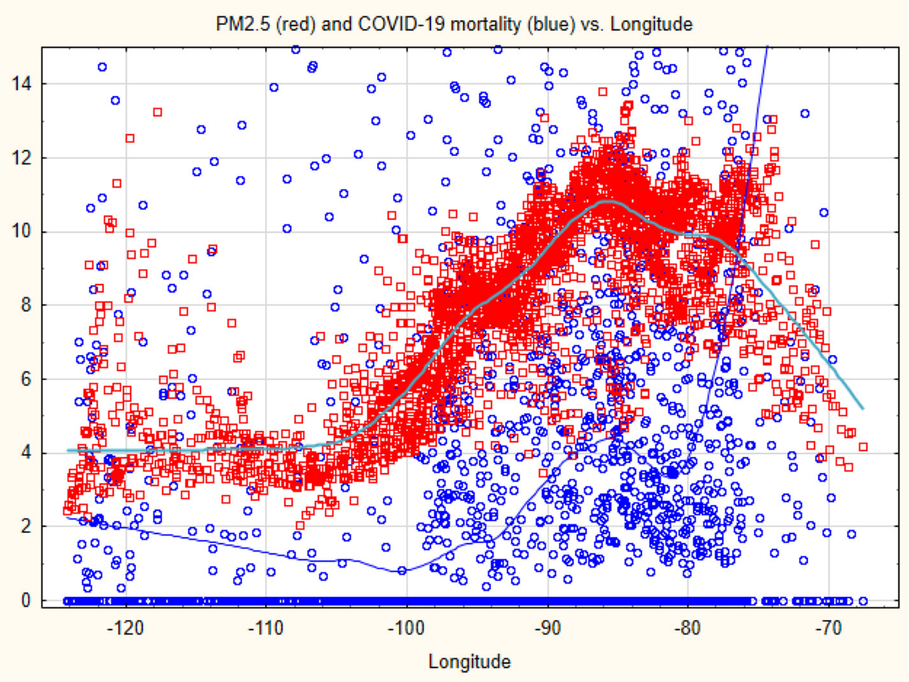


Fig. 5. Scatter plots of average estimated historical PM2.5 concentrations (in micrograms per cubic meter) (red squares) and COVID-19 deaths per 100,000 (blue circles) vs. *Longitude*.

new *Risk* variable defined as $Risk = PopDensityLog^2$. In other words, we artificially create a variable that we know is determined only by population density, via the nonlinear formula $Risk = (\log(\textit{population density}))^2$. (This example is suggested by Fig. 6, which shows a scatter plot of COVID-19 deaths per 100,000 against *PopDensityLog*.) Fitting a multiple linear regression model to the data with this artificial *Risk* variable as the dependent variable yields the results in Table 3. All but one of the predictors, including PM2.5 (*2000-2016AveragePM25*), have highly statistically significant positive regression coefficients, even though, by construction, *Risk* does not depend on anything other than *PopDensityLog*. The reason is that the multiple linear regression model's assumption that risk depends only on a weighted sum of the predictors is false. As illustrated in Fig. 6 for the real risk variable (*DeathRate100k*), risk varies nonlinearly with *PopDensityLog*. The mistaken modeling assumption of linearity is sufficient to induce many other predictors to enter the regression model with significant positive coefficients,

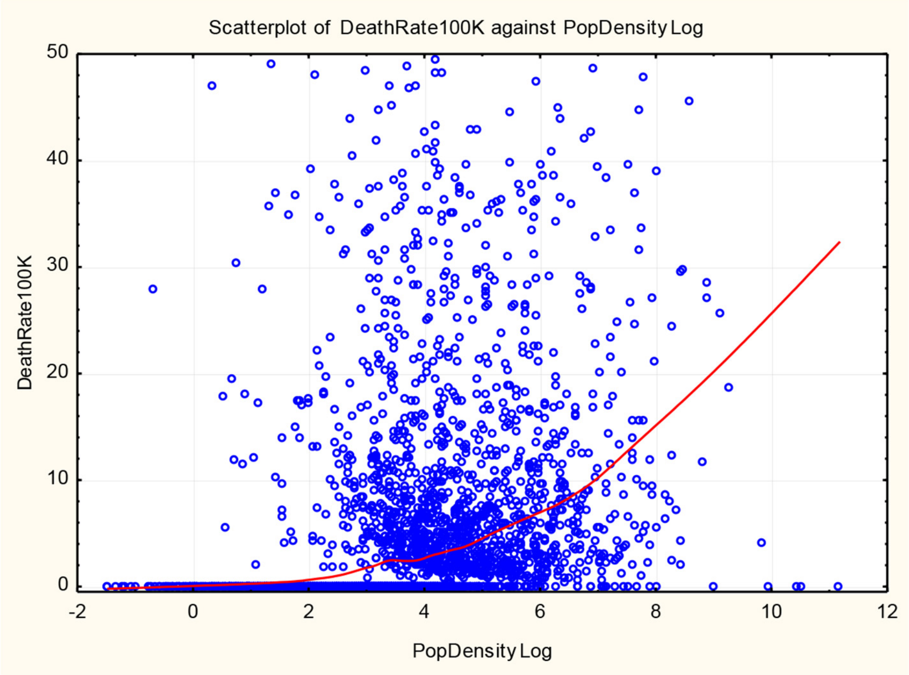


Fig. 6. Scatter plot of COVID-19 deaths per 100,000 (*DeathRate100k*) against *PopDensityLog*. A non-parametric (lowess) smoothing regression curve is superimposed to aid visual interpretation.

Table 3
A multiple linear regression model for the dependent variable *PopDensityLog*^{2}.

$N = 3009$	Regression Summary for Dependent Variable: Risk = PopDensityLog^2 $R = 0.80040234$ $R^2 = 0.641$ Adjusted $R^2 = 0.640$					
	b*	Std. Err.	b	Std. Err.	t (3000)	p-value
Intercept			−47.30	4.19	−11.3	0.000000
2000-2016AveragePM25	0.35	0.02	1.99	0.11	18.9	0.000000
PCT_BLACK	−0.01	0.01	−1.32	1.35	−1.0	0.326789
PCT_HISP	0.16	0.01	17.09	1.44	11.9	0.000000
Latitude	0.37	0.02	1.09	0.07	15.0	0.000000
Longitude	0.17	0.02	0.21	0.02	10.1	0.000000
WINTERAVGTMP	0.22	0.02	0.28	0.03	9.9	0.000000
FirstCaseDays	0.39	0.01	0.33	0.01	33.5	0.000000
PopDensity	0.38	0.01	0.00	0.00	33.1	0.000000

because including them helps to reduce the mean squared prediction error due to model misspecification. Again, interpreting such a positive regression coefficient for exposure as evidence that reducing exposure would reduce risk is mistaken. Instead, positive regression coefficients are only evidence that the assumed regression model does not describe the data.

Nonparametric methods help to avoid these difficulties. Fig. 7 shows a CART tree for the same example as in Table 3. In this tree, as also in a non-parametric Bayesian network fit to the same data, the only predictor of *PopDensityLog*^{2} is found to be *PopDensityLog*.

This example illustrates that even including the right variables (such as measured confounders) in an adjustment set to obtain unbiased estimates of a C-R coefficient in a regression model, controlling for confounders without introducing collider biases [34], does not suffice to prevent spurious significant positive C-R coefficients if the model form is incorrectly specified – for example, by assuming a generalized linear model or no interactions when nonlinearities and interactions among predictors are important, as they are in this example (Figs. 4 and 5). More constructively, it shows that non-parametric methods can help to avoid such spurious C-R coefficients by clarifying which variables provide unique information about a dependent variable (Fig. 7), and which merely reduce errors in predicting the dependent variable by helping to correct for the errors introduced by improper specification of the model (Table 3).

More generally, PM2.5 could have a significant positive regression coefficient as a predictor of COVID-19 mortality risk for any or all of the following reasons [9]):

- Model specification errors, e.g., if mortality rate is assumed to depend on a weighted sum of variables, but in fact its dependence is better described by a model with nonlinearities or interaction terms, as in Table 3.
- Omitted confounders, as in the example of PM2.5 and COVID_19 mortality risk depending on latitude and longitude (independently of other factors, as shown in Fig. 2), if these factors are omitted;
- Measurement errors in explanatory variables, e.g., if PM2.5 is correlated with other variables that are measured or estimated with error, so that including PM2.5 in the regression reduces prediction error due to uncertainties about those variables;
- Residual confounding, e.g., if older people tend to live in more polluted areas and, independently, to have higher mortality rates, but age is only measured in wide categories such as "% of people aged 65 or older";
- Use of surrogate variables, e.g., "Average winter temperature" since 2000, rather than more causally relevant variables such as low temperatures in the months of COVID-19 in 2020;
- Unmodeled interactions or dependencies among variables, e.g., if PM2.5 modifies or is modified by variables such as humidity and temperature that affect respiratory illnesses and COVID-19 mortality;

A positive regression coefficient explained by one or more of these sources does not provide evidence that reducing PM2.5 would reduce mortality risk.

Conclusion: Regression Models and Judgment Should Complement Science, Not Substitute for It

We do not conclude from the foregoing considerations that PM2.5 does not increase risk of COVID-19 mortality; perhaps it does. Rather, we conclude that a positive regression coefficient per se does not provide useful evidence about the matter. This is no straw man argument: as illustrated previously by the examples of BenMAP [29], the claim of Chen et al. [30] that reduced air pollution brought health benefits (based on assumptions rather than observations), and the claim of Wu et al. that "A small increase in long-term exposure to PM2.5 leads to a large increase in the COVID-19 death rate" (based on regression modeling), it remains common practice to present estimated or assumed air pollution concentration-health response (C-R) associations and regression coefficients as if they were known to be manipulative causal relationships. Such regression coefficients are easily produced by modeling choices [33], but lack clear causal interpretation. A judgment that such evidence provides reason to worry – that, in the words of a *New York Times*

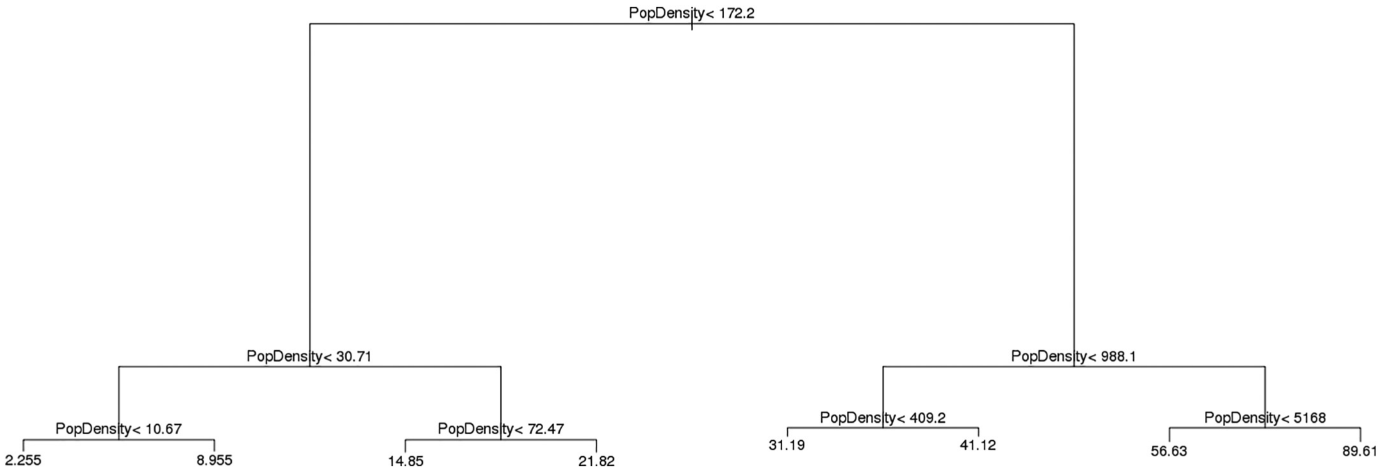


Fig. 7. A CART tree model for the dependent variable *PopDensityLog*^{2}.

headline, "New Research Links Air Pollution to Higher Coronavirus Death Rates" [35] – is simply misleading: the "links" provided by positive regression coefficients are statistical links, not causal ones, and they may signify only that longitude was omitted, or that a linear form was assumed, or that different predictors are correlated with each other and estimated with errors, or that continuous variables have been categorized, and so forth. (All of these are true of the model of Wu et al. [24] behind the headline.) This need not mean that the conclusion is false, but it does mean that the conclusion is not implied by the data and regression analyses from which it is said to be derived. Likewise, attempts to use weight of evidence (WoE) judgments to synthesize all relevant evidence across studies risk producing conclusions of dubious validity if they do not correct such errors and biases in the individual studies being synthesized. Repeating errors many times (as when many investigators fit generalized linear regression models to many different data sets while ignoring key confounders, nonlinearities, interaction terms, etc. in each case) can produce consistency without making results any less erroneous or increasing weight of evidence for a genuine effect.

The analyses presented here illustrate that such errors are easy to avoid using modern data science methods, such as non-parametric trees (or ensembles) to avoid model specification errors and to incorporate nonlinearities and interactions; and tests of conditional independence in BNs (or CART trees) to identify potential confounders that should not be omitted. In the examples presented, these methods show that a statistically significant C-R regression coefficient "linking" PM2.5 to COVID-19 mortality risk could be an artifact of omitted variables and improperly specified parametric regression modeling; for the example in Table 3, this significant positive coefficient disappears when these errors are remedied using non-parametric methods (Fig. 7). Moreover, the diagnosis of which important variables were omitted (such as longitude) and the absence of any detected dependence of COVID-19 mortality risk on PM2.5 once longitude and other variables were conditioned on only required the conditional independence tests built into standard CART tree (Fig. 4) or BN learning software (Fig. 2) – a small part of the arsenal of modern causal analysis techniques. It was not necessary to obtain fully causal BN models with oriented arcs showing the direction of information flow or propagation of changes: the basic criterion that effects should depend upon their direct causes (and hence not be conditionally independent of them) sufficed. The main conclusions – that the data show clearly nonlinear relationships among variables (Fig. 6) and that they do not show a dependence of COVID-19 mortality risk on PM2.5 in conditional independence tests (Figs. 2 and 4) once other important predictors (Fig. 1) have been conditioned on – can be independently verified using the data in Appendix A.

North [7] wrote that "An established paradigm for interpreting epidemiological evidence causally, used by the US EPA, based on considerations proposed in 1965 by British epidemiologist Sir Austin Bradford Hill, is being challenged by another paradigm based on statistical procedures to distinguish between association and causation." Perhaps the most fundamental prescription of the causal paradigm is to recognize that statistical "links" such as positive regression coefficients (or relative risks greater than 1, or positive attributable risks and burdens of disease, and so forth) are neither more nor less than indicators of statistical association, which do not necessarily or usually provide relevant evidence about causation [21]. Modern methods such as the Bayesian network in Fig. 2 can help to discover what evidence a data set does provide that some variables depend, directly or indirectly, on others. For example, Fig. 2 suggests that latitude and longitude have direct effects on COVID-19 mortality risk (meaning, effects in addition to those mediated by the other variables in Fig. 2). This discovery might not have been anticipated intuitively by an investigator, leading to the omission of these confounders, as in Wu et al. [24]. Such computer-assisted discoveries from data may assist, but not replace, the scientific work of formulating testable predictions about whether and how much changes in some variables affect changes in others, and then testing these predictions against new data and reporting the results. If COVID-19 mortality risk appears to be conditionally independent of PM2.5 in non-parametric analyses with adequate power to detect even relatively small effects (Fig. 4), then parametric regression modeling that imposes assumptions on the data sufficient to create a positive regression coefficient for PM2.5 (Tables 2 and 3) should not be construed as evidence that changing PM2.5 would change COVID-19 mortality risk. But neither should it preclude a search for alternative hypotheses, backed by empirical testing, that better explain the observations. For example, for the same average PM2.5 concentration over the past 20 years, do counties with constant or increasing PM2.5 levels over time have significantly earlier first dates of COVID-19 mortalities than counties with PM2.5 levels that decreased over time? Fig. 2 leaves open this possibility, and additional research might pursue it further.

The question of what scientific hypotheses are worth investigating further is surely a proper matter for expert judgment. Interpretation of regression coefficients as evidence that reducing exposure would reduce risk is not. Thus, we conclude that empirical testing of predictive generalizations against data should not be skipped in favor of applying judgment to regression coefficients to draw policy-relevant causal conclusions. Regression coefficients simply do not provide the information needed to determine – or to make sound judgments about –whether or to what extent they are likely to be causal (Table 3). Judgment that seeks to bridge the gap between association and causation based on positive regression coefficients, as in BenMap and its applications [29], is akin to a Rorschach test: an expert may perceive evidence for causation in such coefficients, and may even use them to quantify health benefits to be expected from reducing exposure, but the perceived evidence and expectations of health benefits are solely in the mind of the expert, and neither supported nor refuted by the regression coefficients themselves. The causal paradigm proposes that traditional scientific method, while often more time-consuming and difficult than applying judgment to regression models, is a far more reliable guide to determining whether and to what extent interventions that change exposure will cause risk to change. Verifying scientific models and predictions for PM2.5 and COVID-19 might take impractically long, and meanwhile decisions must be made and risks managed despite scientific uncertainties about causation. But we reiterate that unwarranted causal interpretation of statistical associations and regression coefficients in a WoE framework should not be substituted for sound science. A technically gullible press and policy makers should not be distracted by pre-scientific claims about health effects from PM2.5 based on judgment and regression modeling in the absence of traditional scientific method and careful evaluation of the predictive validity of such claims [27]. Risk analysis and the public interest can and should be better served by adhering to the principles of sound science articulated in the Introduction and to the principles of sound causal inference referred to by North [7].

Declaration of conflict of interest

Dr. Cox is the Chair of the US EPA Clean Air Scientific Advisory Committee. Any opinions expressed are solely those of the authors and not the EPA or any other agencies. No funding was received for this work.

Acknowledgements

We thank five anonymous reviewers and the Editor-in-Chief for their helpful comments and suggestions. These encouraged us to explicitly list how the principles of the scientific paradigm in the introduction can be applied using CART trees, random forest, and Bayesian networks; explain why our critique of regression models is not a universal condemnation of modeling, but only a criticism of clearly identifiable and

avoidable errors such as omitting variables on which exposure and response both depend, or fitting generalized linear main-effects models to data with important nonlinearities and interactions; and clarify that conditional independence tests provide information about evidence of possible causality even if not all arrows in a Bayesian network model have clear causal interpretations. We appreciate the reviewer's questions and comments which contributed to a more explicit discussion of these and other methodological points.

Appendix A. Data and analyses

We collected data from many sources, including most of those cited by Wu et al. (2020), but with alternate authoritative sources for temperature, humidity, and cases/deaths data. We used more recent data for PM2.5, demographics, temperatures, and cases/deaths, and added further sources or fields. For example, we collected USDA county level economic characterizations along with various county attributes compiled by the UC Berkeley Yu Group (2020). Table A summarizes data sources and variables. Data building was accomplished using python scripts. The full data set can be downloaded from http://cox-associates.com/CausalAnalytics/; it is the file “covidpm25.xlsx”.

One option for replicating the random forest, Bayesian Network learning, and CART analyses in this paper is to use the Causal Analytics Toolkit (CAT) software at http://cox-associates.com:8899/. After uploading the data file covidpm25.xlsx (by selecting “Upload File” from the “Data” tab), select the variables to be used in the analysis (e.g., *X2000.2016AveragePM25, PCT_HISP, PCT_BLACK, Latitude, Longitude, PopDensity, PCT_ED_HS, FebMinTmp2000.2019, WINTERAVGTMP, FirstCaseDays* for Fig. 4) and click on “Tree” to generate CART trees (we used the tree generated by the *rpart* package, as this is older and better documented than the *partytree* package). The CAT software provides links to documentation on the algorithms and R packages used; book-length treatments are also available (e.g., Cox LA Jr., Popken DA, Sun RX. *Causal Analytics for Applied Risk Analysis*. Springer, 2018). A short introduction to Bayesian network learning, random forest, and CART algorithms is Cox (2018). In the CAT software, Click on “Importance” to generate random forest importance plots (Fig. 1), and “Bayesian” to generate a BN using the *bnlearn* R package (Fig. 2) [11]. The “Bayesian” option allows constraints to be entered on possible arrow directions. To generate Fig. 2, we specified *Longitude* and *Latitude* as sources (only outward-pointing arrows allowed) and *DeathRate100k* as a sink (only inward-pointing arrows allowed), since latitude and longitude cannot be effects of other variables (but might be causes), and death cannot be a cause of other variables (but might be an effect).

Table A
Data sources and variable overview.

Data Category	Source	Comments
PM2.5	Pm2.5 annual average data from the Atmospheric Composition Analysis Group (http://fizz.phys.dal.ca/~atmos/martin). 0.01° × 0.01° grid resolution PM2.5 prediction in mcg/m^3.	We averaged across grid cells in each county, and produced a 2000–2016 average, as well as separate values for each year 2000–2018.
County boundaries	U. S. Census https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html	Used to provide county boundaries for PM2.5 attribution, land area for popdensity, and centroids lat/lons.
Demographic	U.S. Census Bureau online API. 2018 ACS5 (American Community Survey 5-year data ending in 2018). County level data. https://www.census.gov/data/developers/data--sets/acs-5year.html	List of variables in table below.
Temperatures	NOAA. County level annual data ftp://ftp.ncdc.noaa.gov/pub/data/cirs/climdiv/. Average: climdiv-tmpccy-v1.0.0-20200504' Min: climdiv-tmincy-v1.0.0-20200504' Max: climdiv-tmaxcy-v1.0.0-20200504' Description: county-readme.txt	We averaged across years 2000–2019 for long term averages by month. We also extracted monthly averages for Jan-Apr 2020.
Humidity	Humidity averages by U.S. weather station (city) through 2018. https://www1.ncdc.noaa.gov/pub/data/ccd-data/relhum18.dat. City lat/lons from https://simplemaps.com/data/us-cities.	For each county centroid, the humidity data from the closest (based on lat/lon coordinates) weather station is obtained.
Hospital beds	Hospital level data with county identifier from Homeland Infrastructure Foundation-Level Data (HIFLD) https://hifld-geoplatform.opendata.arcgis.com/-datasets/hospitals as of 10/7/2019.	Aggregated hospitals over counties. Converted to beds per 100 K population. Log version also.
Mitigation policies	State level governmental COVID-19 policies compiled byRaifman et al., Boston University School of Public Health, COVID-19 United States state policy database (www.tinyurl.com/statepolicies).	Used to compute days since stay-at-home order and days since closure of non-essential businesses (from 5/11/2020)
Behavioral	County level data from Robert Wood Johnson https://www.countyhealthrankings.org/ (2020)	Smoking, Obesity, and overall health
County Population	https://www2.census.gov/programs-surveys/popest/datasets/2010-2018/counties/totals/	Used to scale various variables
County attributes	Selected variables from the COVID Severity Forecasting project. UC Berkeley Departments of Statistics, EECS led by Professor Bin Yu https://github.com/Yu-Group/-covid19-severity-prediction. See also https://www.stat.berkeley.edu/~binyu/ps/papers2020/covid19_paper.pdf	List of variables in table below.
Economic characteristics	County level data from USDA - https://www.ers.usda.gov/data-products	3 county coding schemes described in table below.
Outcomes (deaths, cases, days since first case)	Cumulative values by date downloaded from https://github.com/nytimes/covid-19-dataand are as of 5/11/2020	5/11 values for cases and deaths extracted. Converted to per 100 K. Days since first case computed by using first case date.

Table B
Additional variable details.

Variable	Category	Description
PCT_POVERTY	Demographic	% below poverty
PCT_OWNEDHOM	Demographic	% owning home
PCT_ED_HS	Demographic	% with high school education
PCT_BLACK	Demographic	% black
PCT_HISP	Demographic	% hispanic

Table B (*continued*)

Variable	Category	Description
MED_INCOMERATIO	Demographic	Median income, converted to ratio relative to mean over counties
MED_HOMERATIO	Demographic	Median home value, converted to ratio relative to mean over counties
PCT_65PLUS	Demographic	% 65+ years
PCT_45TO64	Demographic	% 45–64 years
PCT_15TO44	Demographic	% 15–44 years
Rural-urban_ContinuumCode_2013	Economic characteristics	1–9 code indicating county degree of urbanization. https://www.ers.usda.gov/data-products/rural-urban--continuum-codes//Created binary column for each level.
Urban_Influence_Code_2013	Economic characteristics	1–12 code indicating county degree of urban influence. https://www.ers.usda.gov/data-products/urban-influence--codes/Created binary column for each level.
Economic_typology_2015	Economic characteristics	1–6 code indicating county economic condition. https://www.ers.usda.gov/data-products/county-typology--codesCreated binary column for each level.
PopDensity[Log]	County Population	2018 Population estimate divided by land area (square miles) from shape files. Log version also.
CensusRegionName	County Attributes	Created binary column for each level.
CensusDivisionName	County Attributes	Created binary column for each level.
StateName	County Attributes	Created binary column for each level.
dem_to_rep_ratio	County Attributes	Ratio of registered democrats to republicans in county
#ICU_beds100K[Log]	County Attributes	Number of ICU beds per 100 K population. Log version also.

References

[1] Cox Jr LA. Should health risks of air pollutants be studied scientifically? Global Epi. 2019;1:100015. https://doi.org/10.1016/j.gloepi.2019.100015.

[2] Craig RT. Constructing theories in communication research. In: Cobley P, Schults PJ, editors. Theories and models of communication. Boston: Walter de Gruyter; 2018.

[3] Cox Jr LA. Improving causal determination. Global Epi. 2019;1:100004. https://doi.org/10.1016/j.gloepi.2019.100004.

[4] Goldman GT, Dominici F. Don't abandon evidence and process on air pollution policy. Science. 2019;2 10.1126;science.aaw9460.

[5] Berrett BT, Samworth RJ. Nonparametric independence testing via mutual information. Biometrika. 2019;106:547–66. https://doi.org/10.1093/biomet/asz024.

[6] Bouezmarni T, Taamouti A. Nonparametric tests for conditional independence using conditional distributions. J Nonparam Stat. 2014;26:697–719. https://doi.org/10.1080/10485252.2014.945447.

[7] North DW. Commentary on “should health risks of air pollution be studied scientifically?” by Louis Anthony Cox, Jr. Global Ep. 2020;2:100021. https://doi.org/10.1016/j.gloepi.2020.100021.

[8] Cox Jr LA. Modernizing the Bradford Hill criteria for assessing causal relationships in observational data. Crit Rev Toxicol. 2018;48:682–712. https://doi.org/10.1080/10408444.2018.1518404.

[9] Cox Jr LA. Implications of nonlinearity, confounding, and interactions for estimating exposure concentration-response functions in quantitative risk analysis. Environ Res. 2020;187:109638.

[10] Vitolo C, Scutari M, Ghalaieny M, Tucker A, Russell A. Modeling air pollution, climate, and health data using Bayesian networks: a case study of the English regions. Earth and space. Science. 2018;5:76–88. https://doi.org/10.1002/2017EA000326.

[11] Nagarajan R, Scutari M, Lebre S. Bayesian networks in R with applications in systems biology. Springer: New York; 2013.

[12] Cox Jr LA, Popken DA. Has reducing fine particulate matter and ozone caused reduced mortality rates in the United States? Ann Epidemiol. 2015;25:162–73. https://doi.org/10.1016/j.annepidem.2014.11.006.

[13] Dorie V, Hill J, Shalit U, Scott M, Cervone D. Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition. Stat Sci. 2017;34(1) 43–68.

[14] Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. Front Genet. 2019;10:524. https://doi.org/10.3389/fgene.2019.00524.

[15] Simon HA. Causal ordering and identifiability. In: Hood WC, Koopmans TC, editors. Studies in econometric method. Cowles Commission for Research in Economics; 1953. p. 49–74 Monograph No. 14. John Wiley & Sons, Inc.: New York.

[16] Dash D, Druzdzel MJ. A note on the correctness of the causal ordering algorithm. Art Intelligence. 2008;172:1800–8.

[17] Heinze-Deml C, Peters J, Meinshausen N. Invariant causal prediction for nonlinear models. https://arxiv.org/pdf/1706.08576.pdf; 2017.

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

[19] Granger CWJ. Investigating causal relations by econometric models and cross-spectral methods. Econometrica. 1969;37:424–38.

[20] Behrendt S, Dimpfl T, Peter FJ, Zimmermann DJ. RTransferEntropy — quantifying information flow between different time series using effective transfer entropy. SoftwareX. 2019;10:100265. https://doi.org/10.1016/j.softx.2019.100265.

[21] Pearl J. Causal inference in statistics: an overview. Stat Surv. 2009;3:96–146.

[22] Triantafillou S, Tsamardinos I. Constraint-based causal discovery from multiple interventions over overlapping variable sets. J Mach Learn Res. 2015;16:2147–205.

[23] Ross SD. The scientific process. Springer Netherlands; 1971 Martinus Nijhoff, The Hague, Netherlands.

[24] Wu X, Nethery RC, Sabath BM, Braun D, Dominici F. Exposure to air pollution and COVID-19 mortality in the United States: A nationwide cross-sectional study. medRxiv. 2020. https://doi.org/10.1101/2020.04.05.20054502 2020.04.05.20054502.

[25] Tollefson J. Air pollution science under siege at US environment agency Top EPA adviser attacks agency decision-making ahead of major review of air pollution standards. Nature. 4 APRIL 2019;568:15–6. https://www.nature.com/articles/d41586-019-00937-w.

[26] Drugmand D. EPA clean air panel chair dismisses his oil industry ties, slams Harvard study on air pollution and COVID risks. https://www.desmogblog.com/2020/05/18/epa-clean-air-committee-cox-harvard-air-pollution-covid.

[27] Burns J, Boogaard H, Polus S, et al. Interventions to reduce ambient air pollution and their effects on health: an abridged Cochrane systematic review. Environ Int. 2020;135:105400. https://doi.org/10.1016/j.envint.2019.105400.

[28] Cox Jr LA. Concentration-response associations used to estimate public health benefits of less pollution are not valid causal predictive models. Ann Am Thorac Soc. 2016;13:2280–1. https://doi.org/10.1513/AnnalsATS.201608-619LE.

[29] Cromar K, Ewart G. Reply: concentration-response associations used to estimate public health benefits of less pollution are not valid causal predictive models. Ann Am Thorac Soc. 2016;13:2281. https://doi.org/10.1513/AnnalsATS.201610-754LE.

[30] Chen K, Wang M, Huang C, Kinney PL, Anastas PT. Air pollution reduction and mortality benefit during the COVID-19 outbreak in China. Lancet Planet Health. 2020. https://doi.org/10.1016/S2542-5196(20)30107-8.

[31] Jiang Y, Wu XJ, Guan YJ. Effect of ambient air pollutants and meteorological variables on COVID-19 incidence. Infect Control Hosp Epidemiol. 2020:1–11. https://doi.org/10.1017/ice.2020.222.

[32] Bashir MF, Ma BJ. Bilal, et al. correlation between environmental pollution indicators and COVID-19 pandemic: a brief study in Californian context. Environ Res. 2020;187:109652. https://doi.org/10.1016/j.envres.2020.109652.

[33] Dominici F, Greenstone M, Sunstein CR. Science and regulation. Particulate matter matters. Science. 2014;344:257–9. https://doi.org/10.1126/science.1247348.

[34] Textor J, van der Zander B, Gilthorpe MS, Liskiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. Int J Epidemiol. 2016 Dec 1;45(6) 1887–1894.

[35] Friedman L. https://www.nytimes.com/2020/04/07/climate/air-pollution-coronavirus-covid.html; 2020.