Commentary

Best Practices for Gauging Evidence of Causality in Air Pollution Epidemiology

Francesca Dominici* and Corwin Zigler

*Correspondence to Dr. Francesca Dominici, Department of Biostatistics, Harvard H.T. Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02115 (e-mail: fdominic@hsph.harvard.edu).

Initially submitted April 18, 2017; accepted for publication August 24, 2017.

The contentious political climate surrounding air pollution regulations has brought some researchers and policymakers to argue that evidence of causality is necessary before implementing more stringent regulations. Recently, investigators in an increasing number of air pollution studies have purported to have used “causal analysis,” generating the impression that studies not explicitly labeled as such are merely “associational” and therefore less rigorous. Using 3 prominent air pollution studies as examples, we review good practices for how to critically evaluate the extent to which an air pollution study provides evidence of causality. We argue that evidence of causality should be gauged by a critical evaluation of design decisions such as 1) what actions or exposure levels are being compared, 2) whether an adequate comparison group was constructed, and 3) how closely these design decisions approximate an idealized randomized study. We argue that air pollution studies that are more scientifically rigorous in terms of the decisions made to approximate a randomized experiment are more likely to provide evidence of causality and should be prioritized among the body of evidence for regulatory review accordingly. Our considerations, although presented in the context of air pollution epidemiology, can be broadly applied to other fields of epidemiology.

air pollution; causal inference; causality; design decisions; study design

Abbreviations: HRS, Huai River Study; ISA, Integrated Science Assessment; SCS, Harvard Six Cities Study; SOSA, Summer Olympic Study in Atlanta; SUTVA, stable unit treatment value assumption; TSP, total suspended particles.

THE RECENT POLITICAL LANDSCAPE SURROUNDING AIR POLLUTION REGULATORY ACTIONS

Evidence from air pollution epidemiology directly informs the development of guidelines and regulations related to air quality in the United States and around the world (1, 2). Because of the contentious political landscape in the United States, air pollution studies have been subject to intense scrutiny.

As part of its review of the current National Ambient Air Quality Standards, the Environmental Protection Agency routinely summarizes the evidence on the harmful effects of air pollution in a document called an Integrated Science Assessment (ISA), designed to provide an overall assessment of the extent to which the relationship between pollution and health can be deemed “causal.” Ultimately, the ISA places this assessment along a continuum according to how likely it is that the relationship between pollution and health can be interpreted as causal, reminiscent of a “classical paradigm” of causality as a degree of evidence (2–5). To inform this judgment, the ISA synthesizes evidence from a variety of studies, including toxicological studies (6), studies that elucidate pathophysiological pathways (7), and epidemiologic studies (2). While the ISA only includes studies that are considered scientifically valid with respect to choice of the study population, study design, confounding adjustment, and reliability of the measurements, the causal validity of each individual study must be evaluated to determine its relative weight in an overall determination of how likely the pollution-health relationship is to be causal. This paradigm has proven invaluable for informing policy decisions related to environmental exposures in the United States and around the world (2–5).

However, notions of what constitutes evidence of causality in air pollution epidemiology are evolving. At least 3 different interpretations exist. First, evidence of causality can be inferred from evidence of biological plausibility as a result of the conduct of toxicology studies with and without randomization (see Stanek et al. (6) for a review on this topic). Second, causality can be determined, as in the aforementioned ISAs, by assessing...
the strength of evidence supporting a particular scientific conclusion. This is based on consistency of results across many epidemiologic studies and adherence to the Bradford Hill criteria (8). Third, “causal inference methods” or “causal analysis techniques” are analytical approaches that, in the context of a single study, are designed to isolate causal relationships. With the recent explosion of important methodological contributions in the area of causal inference methods for observational studies, an increased number of air pollution epidemiologic studies (including our own) have used causal inference data analysis methods with the intent of providing evidence of causality regarding adverse health effects (9–14). Many of these approaches rely on the definition of the counterfactual or potential outcomes, where causality is construed as a comparison between what did happen (e.g., the mortality rate when a population was exposed to high levels of pollution) and what would potentially have happened (possibly contrary to fact—e.g., the mortality rate if a population had actually been exposed to low levels of pollution) (15).

Although it is undeniable that the development of new methods has increased scientific rigor in defining and properly estimating causal effects from observational studies, we argue that the mere application of these methods should not be touted as the silver bullet to finding the truth.

More specifically, increased emphasis on methods for causal inference and their application to the field of air pollution epidemiology is creating a dangerous misconception. Statistical methods for causal inference have played a prominent role in advancing several areas of scientific inquiry (15); see Zigler et al. (9) and Zigler and Dominici (16) for a discussion specific to air pollution epidemiology. However, a false dichotomy is emerging that classifies studies as either causal or associative based solely on the statistical methods used for estimation. Studies that estimate exposure-response relationships with regression models typically do not use causal inference terminology and are consequently coined associative and considered less rigorous. Conversely, studies that use explicit causal inference terminology are labeled causal, which automatically endows them as a fail-safe solution for identifying the scientific truth. This misconception is dangerous in the context of environmental regulations and spreads a false message that studies using causal inference methods should always be considered more credible than studies using more traditional statistical approaches.

In this commentary, we argue that assessing the causal validity of an air pollution study should focus not exclusively on the statistical methods used for estimation but most importantly on the design decisions that render the analysis of observational data an approximation of the analysis of a randomized experiment (17).

As we elaborate below, the term design decisions as used in this commentary entails more than just traditional notions of “study design” (e.g., whether the study is a case-control study, prospective cohort study, or case-crossover study). It also refers to analytical decisions about how to make effective use of data (e.g., how to construct a good comparison group) arising from any number of data-collection designs.

We argue that by critically evaluating these design decisions, we can determine how well they approximate an idealized randomized study of the population of interest, with closer approximations providing more support for causal validity. We use the term idealized randomized study because the randomized studies that are routinely conducted have very strict inclusion criteria, which are deemed necessary to ensure the internal validity of the study at the expense of external validity. To be informative to the regulatory process, external validity is necessary. Thus, our objective is to examine design decisions that will allow researchers to approximate an idealized randomized study that will include the population of interest and estimate the desired causal effects.

THREE AIR POLLUTION STUDIES AS EXAMPLES

We illustrate the idea of design decisions in the context of 3 air pollution studies: 1) the Huai River Study (HRS) (18), a quasi-experimental study; 2) the Summer Olympic Study in Atlanta (SOSA) (19, 20), an intervention study; and 3) the Harvard Six Cities Study (SCS) (21), a cohort study. The HRS is a study of an open cohort analyzed as a cross-sectional study; the SOSA is a study of an open cohort analyzed as a time series; and the SCS is a study of a closed cohort analyzed as a survival study.

The Huai River Study

In 2013, Chen et al. (18) conducted a study which leveraged a government policy in China that provided free coal for indoor heating north of the Huai River (which in turn led to high concentrations of total suspended particles (TSP)) but not south of the river between the 1950s and the 1980s. The provision of this policy led to a quasi-experimental approach for estimating the causal effect of exposure to TSP on life expectancy. Because of the policy, levels of TSP were much higher north of the river than south of the river, but other potential confounders (socioeconomic status, smoking, comorbidity) did not differ between the two areas.

The HRS is quasi-experimental under the assumption that exposure to high levels of TSP versus low levels is governed by an arbitrary policy that is unrelated to life expectancy. By fitting 2 regression models, the authors separately tested whether the Huai River policy caused a discontinuous change in TSP levels above the river and a discontinuous change in life expectancy. They wrote, “[The] availability of a regression discontinuity design based on the Huai River policy provides an appealing quasi-experimental approach that can help to move the existing literature from documenting a robust association between particulates and health toward documenting a causal relationship” (18, p. 12936). The regression discontinuity design was developed more than 5 decades ago and has been used successfully to test the causal nature of relationships in a wide range of fields, including psychology, education, statistics, biostatistics, and economics (22, 23).

The Summer Olympic Study in Atlanta

Peel et al. (20) conducted a study to assess the effects of a traffic ban imposed in Atlanta, Georgia, during the 1996 Summer Olympic Games (July 19, 1996–August 4, 1996) on asthma-related hospital emergency visits, but this study is not labeled as causal. The investigators gathered daily data on numbers of emergency department visits, air pollution levels, traffic levels,
and several meteorological variables across 5 counties in the Atlanta area for the summer months just before, during, and after the Olympic Games (1995–2004). They compared daily pollutant concentrations and traffic counts between the Olympic and baseline periods using a time-series approach. To estimate the effect of the intervention on emergency department visits, Peel et al. used Poisson generalized linear models of the daily number of emergency department visits during the Olympic and baseline periods in the years 1995 through 2004. The analyses were adjusted for day of the week, daily minimum temperature, daily average dew point temperature, day of summer, and year.

The Harvard Six Cities Study

Dockery et al. (21) conducted a prospective cohort mortality study involving 8,111 randomly selected residents of 6 US cities (Harriman, Tennessee; Portage, Wisconsin; St. Louis, Missouri; Steubenville, Ohio; Topeka, Kansas; and Watertown, Massachusetts). The researchers assembled data on total suspended particulate matter, sulfur dioxide, nitrogen dioxide, and ozone, among other factors, for these 6 locations. Based on more than a dozen years of exposure data provided by ambient air monitoring stations, Portage and Topeka had the lowest concentrations of particulate matter pollution, whereas Steubenville and Harriman had the highest concentrations. The researchers’ objective was to estimate the effects of long-term exposure to air pollution on mortality while controlling for other risk factors, such as individuals’ smoking status and age. Residents of the 6 cities were recruited between 1974 and 1977, completed questionnaires about their medical and occupational history, and underwent lung function (spirometry) tests. From 1974 to 1991, the researchers contacted the residents annually to determine their vital status. They analyzed the data by fitting a Cox proportional hazards model and adjusted for individual-level risk factors and potential confounders by including these variables in the Cox regression model. The SCS has been labeled an associational study because there is no specific language regarding causality in the published article (21) and data were analyzed within a regression framework.

BEST PRACTICES FOR DETERMINING EVIDENCE OF CAUSALITY IN AIR POLLUTION EPIDEMIOLOGY

Our proposed best practices for determining the degree to which an air pollution study can provide evidence of causality are not new and indeed have been described in the epidemiologic literature numerous times (e.g., see references 24–27). In this commentary, we review some of these best practices and illustrate them specifically in the context of the case studies cited above. Our focus here is on evaluating the threat of confounding bias, which is the most commonly cited threat to the causal validity of air pollution studies, although in general other biases may also be important and may require additional methods (28–30). Our considerations are not necessarily tied to the specific statistical method used for the data analysis. Instead, we argue that interrogating a study to discern how closely design decisions approximate an idealized randomized experiment plays a far more important role than simply looking at whether the methods used for estimation are labeled as causal.

Identify the experimental conditions and the potential outcomes

The first step for evaluating causal validity is identifying the conditions (actions or exposure levels) that created the comparison of interest. In order to determine whether one of the conditions (“action A”) caused an outcome, one must infer what the potential outcome would have been under the alternative condition (“action B”). In the HRS, determining whether high levels of TSP north of the Huai River (action B) caused shorter life expectancy requires knowing what the mortality rate would have been if TSP levels north of the river remained as low as those in the south (action A). In the SOSA, concluding that the traffic ban (action B) caused a reduction in asthma visits requires knowing that there would have been more asthma visits if the traffic ban had never been implemented (action A). The SCS was purposely designed to compare 6 actions: the observed air pollution levels in 6 US cities. Determining whether the levels in one city, say Steubenville (action B), caused mortality requires knowing what the mortality rate would have been if the people of Steubenville had been exposed to lower levels, say those of Portage (action A).

Define the causal effect of interest

A clear definition of the conditions and potential outcomes, or a comparison between what was observed to happen and what would have happened under alternative conditions, leads to a clear definition of the causal effect. In the HRS, the causal effect was the effect of being exposed to high TSP levels north of the river versus lower levels south of the river on life expectancy. In the SOSA, it was the effect of experiencing a traffic ban versus not experiencing a traffic ban on asthma visits. In the SCS, it was the effect of being exposed to high levels of air pollution in Steubenville versus lower levels in Portage on mortality. The fundamental challenge is that we can never observe a potential outcome under a condition other than the one that actually occurred. Thus, causal inference relies on the ability to predict the unobserved potential outcome.

Understand the threat of confounding

If an experiment were conducted to randomly assign observations to actions A and B, then prediction of the potential outcomes would be straightforward. The goal of the randomization is to ensure exchangeability—in other words, to ensure that there are no systematic differences between persons exposed to action A and those exposed to action B, rendering average outcomes among those exposed to action A a reliable prediction of what would have happened to those exposed to action B (and vice versa). Without randomization, study units that receive action A versus action B could be systematically different and therefore not exchangeable. Confounding occurs when the distributions of covariates that affect both the treatment assignment and the outcome differ between the two groups. In the HRS, people living north of the river may spend more time indoors than people living south of the river, and spending time indoors could be associated with life expectancy. In the SOSA, the timing of the traffic ban may have coincided with other climatological or behavioral changes that would have occurred regardless
of the ban, and these climatological or behavioral changes might affect asthma admissions. In the SCS, persons exposed to higher air pollution levels in Steubenville may have had different individual characteristics (e.g., lower income) than those exposed to lower levels in Portage, and lower income is associated with higher mortality risk. The threat of confounding is that differences in the outcomes when comparing action A with action B could instead be due to differences between the experimental units exposed to these two actions.

**Construct an adequate comparison group**

The issue of confounding can be resolved by constructing an adequate comparison group: finding observations that are exactly the same except for their exposure to action A versus action B, as would be the case if observations were randomized to action A versus action B. In the HRS, the investigators used a regression discontinuity design (18), which relies on the assumption that the arbitrary boundary of the heating policy was the only determinant of TSP exposure levels and that this was the only systematic difference between populations living just north and just south of the river. This assumption renders those living just south of the river a reliable comparison group for those living just to the north. In the SOSA (see Figure 1), Friedman et al. (19) constructed a comparison group from the same population but a different time period, using data from the 4 weeks before and after the Olympic period as a comparison group for the Olympic period to adjust for trends (black lines (condition A) in Figure 1). This relies on the assumption that other predictors of asthma (e.g., weather) present during the comparison period were also present during the Olympic period. To better characterize what ordinary seasonal variation in asthma visits would have been without the traffic ban, Peel et al. (20) conducted follow-up analyses for the years 1995 and 1997–2007 during the summer weeks of the Olympic period and the 4 weeks before and after the Olympic period (red lines (condition A′) in Figure 1). In the SCS, regression adjustment was implemented with the intent of ensuring that the residents of Steubenville were compared with a comparison group consisting of Portage residents with similar confounding characteristics (and analogously for other cities).

We argue that a key consideration in determining whether a statistical analysis supports causality is whether it relies on data from adequately constructed comparison groups. Ideally, one would construct a comparison group (or period) by finding, for every observation exposed to action B, an observation that was exactly the same (based on confounders) but exposed to action A. Figure 2 schematically represents the idea of constructing a comparison group in the SCS by exact matching on measured confounders. For example, say that Jane, a participant in the SCS, lives in Steubenville, does not smoke, and has a low socioeconomic status. A comparison group would be constructed by matching Jane to a group of women with the same confounding characteristics (the same age, smoking status, and socioeconomic status) who live in Portage (action A). The comparison group is represented by A′ in Figure 2. If all relevant confounders are comparable, the matched Portage residents could represent the unobserved potential outcome for Jane had she been exposed to the levels of particulate matter in Portage. Averaging such matched comparisons across a study population could estimate causal effects.

When the number of confounders is very large, matching all confounders exactly might not be feasible. A common approach to addressing this challenge is to use propensity scores. First, estimate the propensity score to represent a one-number summary of all confounders and then use this score to construct a comparison group by either matching, stratifying, or weighting observations based on the propensity score (31).

---

**Figure 1.** Time-series plots of daily numbers of asthma-related emergency department visits to 12 hospitals in 5 counties in Atlanta, Georgia, for the summer Olympic period in 1996 and for the average of the years 1995 and 1997–2004 (comparison group). “A” denotes the daily number of asthma visits for the days before and after the Olympic period (black time series); “B” denotes the daily number of asthma visits during the Olympic period (blue time series). A’ denotes the daily number of asthma visits for the comparison group. Reproduced from Peel et al. (20) with permission (slightly modified).
causal effect could be biased. Often, circumstances generating quasi-experiments, such as those in the HRS, are leveraged to automatically adjust for both observed and unobserved confounders, which holds as long as the untestable assumption that treatment assignment does not depend on any determinant of the outcome is valid (34–36).

Additional assumptions to identify a causal effect. As reviewed by Hernán (34), there are many additional threats to the validity of being able to identify a causal effect in addition to the one of no unmeasured confounders. They include 1) selection bias, 2) positivity, 3) the stable unit treatment value assumption (SUTVA), and 4) confounding by coexposures. Selection bias is caused by preferential exclusion of samples from the data (e.g., nonignorable dropout) and is a major obstacle to valid causal and statistical inferences (34–36). Positivity is the assumption that any experimental unit has a positive probability of receiving all values of the exposure variable. SUTVA requires that the potential outcome on one unit should be unaffected by the particular assignment of treatments to the other units (37). Violations of SUTVA can prove particularly challenging in studies of air pollution interventions, as the transport of pollution through space often presents settings of interference. Intervening at one location (e.g., a pollution source) likely affects pollution and health across many locations, meaning that potential outcomes at a given population location are probably functions not just of local interventions but also of interventions at other locations (38, 39). In air pollution epidemiology, often we are dealing with the issue of confounding due to highly correlated coexposures, such as those in SOSA. To best inform regulatory policy, which regulates one pollutant at a time, it is often of interest to estimate the causal effect of exposure to 1 pollutant (e.g., fine particulate matter) adjusted for coexposure to another pollutant (e.g., ozone). In this context, use of a proper study design to minimize the correlation between coexposures is again essential for evidence of causation by a specific exposure. Another important issue to consider is the interpretation of marginal effect versus conditional effect: A nice feature of causal inference methods is that they estimate a “marginal effect”—that is, the average effect at the population level—of moving an entire population from untreated to treated, whereas more traditional estimation approaches, such as regression models, random-effect models, multistage models, etc., commonly estimate a “conditional effect”—that is the effect of a change in exposure on the outcome with the values of all the other covariates kept constant (40, 41). A detailed discussion of each of these assumptions in the context of air pollution epidemiology is beyond the scope of this paper.

CONCLUSIONS

In this commentary, we have attempted to improve clarity and transparency regarding the most important tenets for discerning whether an air pollution study provides evidence of causality. Our intent is by no means to disparage the use of causal inference methods. On the contrary, we merely hope to identify the false dichotomy created by the assumption that statistical analyses explicitly labeled as “causal inference” or “causal analysis” methods are, by default, superior. A vast array of analytical approaches may not be specifically labeled as “causal”

Perhaps the most common method of adjusting for observed confounding is to fit a regression model that includes the indicator of being assigned to A versus B and the potential confounders as covariates. Such an approach is similarly motivated by the desire to construct an adequate comparison group, but rather than explicitly matching on covariates or using the propensity score, regression modeling relies on parametric assumptions to model what would have potentially happened to an observation under an alternative exposure. Regression is frequently combined with strategies based on propensity score (or other forms of) matching or weighting, often in the form of doubly robust estimators (32, 33). In lieu of constructing comparison groups based on measured confounders, quasi-experiments instead make the assumption that no confounding adjustment is necessary due to the specialized features of the design. Please note that regression adjustment in the SCS highlights the distinction between design decisions and analysis decisions. Using 6 experimental conditions is a design decision for comparing survival experiences across different levels of air pollution. The estimation of the exposure-response curve is an analysis decision: The exposure levels in the 6 cities are used to inform a regression model representing the (confounding-adjusted) relationship between changes in air pollution and health outcomes.

The threat of unmeasured confounding. If there were important confounders that were unmeasured and uncorrelated with all of the measured confounders, then the study design would poorly approximate randomization and the estimated average
or explicitly rooted in potential outcomes but are nonetheless implemented to alleviate the same threats to validity that can be addressed with rigorous and careful design decisions (37, 42). Clearer understanding of this false dichotomy will help to refocus the scientific debate around causality and encourage a closer examination of design decisions, rather than simply concentrating on nominal distinctions between “causal” and “associational” statistical methods.

We argue that a critical assessment of the design decisions is essential for judging causality, in some instances more so than the degree of sophistication of the statistical methods used for estimation, especially in the context where the sophistication of the statistical methods is not directly targeting the shortcomings of the design. Different studies with their own design considerations support different degrees of causal validity, but we argue that this distinction is not simply about the statistical estimation procedure used. In summary, best practices for assessing the degree to which an air pollution study provides evidence of a cause-and-effect relationship should focus on addressing the following questions:

- What type of design was used?
- What are the actions or exposure levels being compared?
- What are the potential outcomes being compared?
- What type of design was used?
- What is an adequate comparison group or comparison period?
- How is the comparison group/period used to estimate the unobserved potential outcomes?
- Why might the chosen comparison group/period not accurately represent the unobserved potential outcomes?
- To what extent does the statistical analysis overcome the limitations of the design decisions?

We argue that air pollution studies that are more scientifically rigorous in terms of the decisions made to approximate an idealized randomized study design are more likely to provide evidence of causality and therefore should play a more predominant role in the regulatory setting.

ACKNOWLEDGMENTS

Author affiliations: Department of Biostatistics, Harvard T. Chan School of Public Health, Boston, Massachusetts (Francesca Dominici, Corwin Zigler).

This work was made possible by support from National Institutes of Health grants R01 ES024332-01A1, ES-000002, ES024012, and R01ES026217; National Institutes of Health/National Cancer Institute grant R35CA197449; Health Effects Institute grant 4953-RFA14-3/16-4; and Environmental Protection Agency grants 83587201-0 and RD-83479801.

This paper originated as a short course presented to an audience of environmental epidemiologists, statisticians, policymakers, and industry consultants. The workshop was offered to approximately 100 people as part of the 2015 conference of the Health Effects Institute (Boston, Massachusetts).

The contents of this article are solely the responsibility of the grantee and do not necessarily represent the official views of the funding agencies. Further, the funding agencies do not endorse the purchase of any commercial products or services related to this publication. Conflict of interest: none declared.

REFERENCES


