Accountability Assessment: An Emerging Opportunity for Causal Inference in Air Pollution Epidemiology

Corwin M. Zigler and Francesca Dominici
Department of Biostatistics, Harvard School of Public Health

1 Introduction: Paradigms for Causal Inference in Epidemiology

The motivation for establishing causal relationships in air pollution epidemiology is clear; interventions (e.g., regulatory actions) are based on presumed causal relationships between exposure to air pollution and health. The challenge is to generate causal evidence from available observational data. Causal inference in air pollution epidemiology typically relies on determination of whether estimated associations between exposure to air pollution and disease can be reliably interpreted as causal relationships (Dockery et al., 1993; Laden et al., 2006; Zeger et al., 2008; Pope III et al., 2009; Correia et al., 2013). This constitutes a so-called “classical” paradigm of causality (Glass et al., 2013), where associations from epidemiological studies (e.g., an exposure-response curve) are carefully examined and evidence is synthesized across studies to determine the extent to which observed relationships can be deemed causal. In this paradigm, interpretation of associations as causal is based on scientific criteria such as those postulated by Sir Austin Bradford Hill in the debate over cigarette smoking in the 1950s and 1960s (Hill, 1965). Often, evidence is considered on a continuum, with more rigorously-generated evidence considered more likely to be causal. The classical paradigm has proven invaluable for informing policy decisions pertaining to environmental exposures (U.S. EPA, 2009; IARC, 2006; National Research Council, 2005; Institute of Medicine, 2008), but establishment of causality under this paradigm does not rely on the definition of any specific action or intervention.

Accountability assessment is a unique area of air pollution epidemiology for its direct connection to specific actions or interventions. This presents an opportunity to incorporate a different causal paradigm into the field: a potential-outcomes paradigm for causal inference rooted in the core tenets of experimentation (Neyman, 1923; Rubin, 1978; Holland, 1986; Hernan et al., 2008). This paradigm explicitly defines causal effects as consequences of specific interventions, lending increased precision to the notion of cause and effect. This is an important departure from the classical paradigm of causality designed to establish and interpret evidence of associations between risk factors and health outcomes. The purpose of this paper is to distinguish between the classical and potential-outcomes paradigms for causal inference as they relate to accountability assessment and to advocate increased use of statistical methods rooted in a potential outcomes framework for causal inference (Rubin, 1978). Accordingly, we highlight several challenges that deserve consideration in conducting accountability assessment of past and future air quality interventions.

1.1 The Classical Framework for Causal Inference

The classical framework for causal inference aims to characterize risk factors (e.g., air pollutants) that exhibit causal associations with health outcomes. This paradigm is evident in EPA Integrated Science Assessments that classify strength of available evidence as “not likely,” “inadequate,” “suggestive,” “likely,” or “causal” (U.S. EPA, 2009; IARC, 2006; National Research Council, 2005; Institute of Medicine, 2008). This classification scheme provides a structured approach to evaluating the evidence for causal relationships, allowing for informed decision-making regarding regulatory actions and public health interventions.
Viewing evidence on a continuum reflecting how likely observed associations are causal lacks any precise notion of cause and effect and is subject to a certain degree of subjectivity that can lead stakeholders with opposing views to have different interpretations of the available evidence (Glass et al., 2013). Once causal validity of associations is established, incorporating these relationships into policy decisions entails an implicit belief that observed relationships would persist upon future interventions to impact air quality.

1.2 The Potential Outcomes Framework for Causal Inference

Randomized controlled experiments are generally accepted as the “gold standard” for generating causal evidence. The potential-outcomes framework for causal inference frames observational studies of environmental interventions on a continuum according to how well they can approximate a randomized experiment (Rubin, 2008a). In contrast to the classical paradigm, the potential-outcomes paradigm involves explicit definition of an intervention that can be conceived of as having arisen from a hypothetical experiment with (at least) two intervention conditions (e.g., enacting a regulatory intervention vs. no regulation). Causal effects are then explicitly defined as comparisons between outcomes under different intervention conditions, which is formalized with the notion of a potential outcome denoting what would have potentially happened if an alternative intervention had been implemented but everything that was not a consequence of the intervention remained the same (Rubin, 1978). Thus, the salient question for accountability assessment is not “Did air pollution and health outcomes change after the intervention?”, but rather “Are air quality and health outcomes different after the intervention than they would have been if some other action had been taken?” (see Figure 1). In a potential-outcomes framework, clear specification of an intervention yields clear definition of causal effects as consequences of specific actions. This framework also appears in Integrated Science Assessments that consider distributions of health outcomes under various possible intervention scenarios (including no intervention) (U.S. EPA, 2009).

1.3 Accountability Studies as Approximate Randomized Experiments

Making causal inferences under the potential-outcomes paradigm relies on two essential tasks (Rubin, 2008a). The first is to frame accountability assessment as a hypothetical experiment consisting of (at least) two intervention conditions, where one condition is often akin to a “placebo” condition where no action is taken. Because causal effects are ultimately defined as outcome comparisons under the different intervention conditions, explicit definition of the hypothetical experiment formalizes the research question at hand. The definition of the experiment determines the definition of the causal effect. Importantly, defining the research question via definition of the causal effect is carried out without regard to any assumed statistical model, and is typically done before seeing any data.

The second task of the potential-outcomes paradigm is formulation of the assignment mechanism that reflects exactly how observational units (e.g., pollution monitor locations, counties, zip codes, individuals) were “assigned” to their respective intervention condition (Rubin, 1991). For example, a regulatory action may apply to some areas but not others, and may be instituted at a single time point or occur in phases. After formalization of the intervention and the assignment mechanism, the fundamental problem is that any particular area can only be subject to one intervention condition, which requires prediction of the unobserved potential outcomes that would have occurred under a different intervention condition. This task may (although may not necessarily) rely on models for pollution and health outcomes.

Existing accountability studies have been conducted under this paradigm, often through framing as “natural experiments.” The benefit of leveraging natural experiments for accountability assessment is that they often permit standard epidemiological tools to yield causal conclusions. For example, many existing accountability studies rely on specific, localized actions that lead to abrupt and dramatic changes in air quality and/or health outcomes.
Causal analyses of these studies can be relatively straightforward because the assumptions required to interpret temporal trends in pollution and/or health as causal effects of the intervention are somewhat uncontroversial. Substantial and immediate changes in pollution leave little ambiguity as to whether these changes are in fact caused by the actions under question.

To fix ideas, consider one flagship illustration of accountability, namely, the analysis of the effects of banning the sale, marketing, and distribution of coal in Dublin, Ireland (Clancy et al., 2002). The coal ban represents a specific localized action (the ban) that was followed by significant decreases in the concentration of black smoke immediately following the ban, with concurrent decreases in mortality that persisted for at least 72 months. Interpretation of these changes in pollution and health outcomes as causal effects of the ban within a potential-outcomes paradigm is fairly straightforward. The implicit hypothetical experiment is one that compares implementation of the ban against an alternative intervention condition where no such ban is instituted. This defines causal effects as comparisons between potential outcomes under the ban (which are observed) and potential outcomes that would have occurred in the absence of the ban (which are unobserved). The ban was “assigned” at a single point in time across the entire city. They key assumption underlying interpretation of post-ban changes in pollution and health as causal effects of the ban is that of temporal stability (Holland, 1986) which assumes that black smoke and mortality would not have changed during this time frame without the ban. Temporal stability in this case assumes that the potential outcomes under the hypothetical intervention condition where no ban was implemented would resemble those that were observed immediately preceding the ban. This permits interpretation of post-ban changes in pollution and health as causal effects of the ban (see Figure 1(a)).

Unlike the coal ban, complex regulatory interventions (e.g., those emanating from the Clean Air Act) are likely to produce subtle effects on air pollution and health over an extended time frame. This presents many challenges to causal inference. For example, consider a proposed accountability assessment of the initial PM$_{10}$ nonattainment designations following the 1990 Clean Air Act Amendments. The nonattainment designations do not represent any single localized action to control air quality. The heterogeneity of actions taken according to State Implementation Plans responding to nonattainment designations complicates even the formulation of the accountability question, as the diversity of actions does not immediately indicate a single hypothetical experiment. This difficulty is in addition to the challenge of attributing evident changes in pollution and health during the 1990s to the designations themselves. Assumptions about what would have occurred in the absence of the nonattainment designations (e.g., temporal stability) become increasingly tenuous when considering longer time frames between promulgation of regulations and noticeable changes in pollution and health. As a result, analyses of temporal trends in pollution and health are inappropriate for identifying causal effects of broad and complex regulatory actions (see Figure 1(b)). Are improvements in air quality due to the nonattainment designations or due to other features of the Clean Air Act enacted during a similar time frame? Are improvements in health outcomes attributable to reductions in air quality, or are they simply consequences of advances in medical care that have been achieved over a similar time frame? These inherent difficulties in conducting long-term accountability assessment of broad regulatory actions require analysis tools specifically tailored to estimation of causal effects. This is perhaps why long-term accountability assessments of broad regulatory strategies are relatively uncommon (Chay et al., 2003; Zigler et al., 2012; Deschenes et al., 2012).

2 Methods for Causal Accountability Assessment

Modern methods for causal inference rooted in a potential outcomes framework span literature in a variety of areas such as statistics, epidemiology, econometrics, political science, implementation science, and program evaluation. Common strategies used within this paradigm include propensity score methods (Rosenbaum and Rubin, 1983), regression discontinuity designs (Imbens and Lemieux, 2008), marginal structural models (Robins et al., 2000),
instrumental variables analyses (Angrist et al., 1996), and many others (useful reviews appear in Rubin (2008b); Imbens and Wooldridge (2008), and Greenstone and Gayer (2009)). The following outlines some of the most salient challenges for incorporating these methods into accountability assessment. While we illustrate concepts with a proposed accountability assessment of the initial PM$_{10}$ nonattainment designations following the 1990 Clean Air Act Amendments, strategies to collect data to address these challenges will be integral for the planning of future accountability assessments.

2.1 Defining the Intervention

Defining the intervention (i.e., “the cause”) in the context of a hypothetical experiment focuses the research question at hand and defines causal effects of interest. Importantly, this task is undertaken without regard to any specific statistical model, and is typically done before seeing any data. While definition of the intervention is often straightforward, as in the coal ban example, it can be subtle for complex regulatory policies.

Consider the PM$_{10}$ nonattainment designations. One natural approach is to frame accountability assessment as evaluation of a two-armed experiment, with one arm consisting of areas designated as nonattainment (the “active treatment” arm) and another arm consisting of attainment areas (the “control” arm). Framing accountability assessment of the PM$_{10}$ nonattainment designations in this way implies definition of the causal effect of the nonattainment designations themselves and not, for example, the effect of actual control measures. Thus, the causal effect is unambiguously defined as the comparison between what happened under the nonattainment designations to what would have happened absent the designations, but this may or may not correspond to specific actions taken in a State Implementation Plan. As with causal assessment of the PM$_{10}$ nonattainment designations, the increased precision in definition of causal effects of regulatory policies often comes at the sacrifice of some level of specificity with regard to actual control measures. In many cases, it may prove valuable to evaluate the causal effects of regulatory decisions (e.g., nonattainment designations), even when these decisions can lead to a variety of actions. Key to the design of future accountability assessments is the availability of data to characterize the intervention with the desired level of specificity. If interest lies in specific actions taken as a consequence of a regulatory decision, the estimation of causal effects of these actions requires detailed data on when and where these actions were taken.

2.2 Confounding and Control Populations

One motivation for framing accountability studies as hypothetical experiments is that this paradigm forces consideration of an appropriate control population that can approximate what would have happened in the absence of the intervention being studied. The control population could be defined based on time (e.g., Dublin before the coal ban can be a control population for Dublin after the coal ban) or could be comprised of different areas or units (e.g., attainment areas could be a control population for nonattainment areas). Consideration of the assignment mechanism aids identification of an appropriate control population.

The key issue with identifying a control population pertains to how comparable it is to the population upon which the intervention is enacted. The notion of “comparability” boils down to the familiar concept of confounding, although the notion of a “confounder” is slightly different under the classical and potential-outcomes paradigms. In the classical paradigm, confounders are generally considered to be anything associated with both exposure to pollution and health outcomes. In contrast, the potential-outcomes paradigm defines a confounder as any factor that differs between intervention groups and also has some bearing on the outcome(s) of interest. Randomized experiments ensure that the treatment groups do not differ from one another (i.e., are comparable), meaning that randomized experiments are “unconfounded.” Observational studies do not automatically enjoy the benefits of unconfoundedness (also referred to as “ignorable treatment assignment”) (Rubin 2008a). If intervention groups in
an accountability study differ on factors related to health or pollution outcomes, then the hypothetical experiment is confounded. For example, if areas subject to an intervention tend to be more densely populated than control areas and population density is directly or indirectly related to pollution or health outcomes, then population density is a confounder. If, however, intervention groups are comparable with respect to population density, then population density may not be a confounder despite its relationship with pollution and health.

Practically speaking, the notion of a confounder in the potential-outcomes paradigm is essentially the same as the classical definition as a factor that is associated with both exposure and outcome. The key difference is that, in the potential-outcomes paradigm, the “exposure” is the intervention and pollution and health measures are outcomes. Confounding adjustment in the potential-outcomes paradigm consists of methods that can adjust for factors that differ between intervention groups. If the intervention and control groups can be “adjusted” so that they appear similar on the basis of all necessary confounders (e.g., through matching, weighting, stratification, or standardization), then the hypothetical intervention can be construed as an approximate randomized experiment. Within groups of units having similar confounder distributions, “assignment” to the intervention is random in the sense that it is unrelated to pollution and health. This type of confounding adjustment frequently appears in the form of the propensity score [Rosenbaum and Rubin 1983], which underlies many methods for causal inference. However, this can be particularly challenging for accountability assessment of air quality regulations, as interventions are typically enacted in areas with high pollution, which may indicate important differences with areas that were not subject to the intervention.

Consider again the PM\textsubscript{10} nonattainment example. The challenge is to predict what would have happened to air pollution and health in nonattainment areas had these areas not been designated nonattainment. Attainment areas present one natural choice for a control population. The success of using information in attainment areas to learn about what would have happened in the nonattainment areas hinges on ability to adjust for confounders such that attainment areas can be considered comparable to nonattainment areas. However, some nonattainment areas simply do not have comparable attainment areas, even after confounding adjustment. For example, Los Angeles County, which was designated as nonattainment for PM\textsubscript{10}, is unique with regard to its demographic makeup and pollution levels. Finding attainment areas that can be considered representative of Los Angeles County is difficult, if not impossible.

The inability to find representative control units for some intervention areas simply prevents estimation of a causal effect for those areas without statistical models that extrapolate relationships beyond the distribution of observed data. Often, intervention areas with no comparable control population are omitted from causal analyses because the data provide no sound basis to make a causal conclusion [King and Zeng 2006; Ho et al. 2007]. If researchers wish to extrapolate beyond the available data to make causal conclusions, then the experimental paradigm should, at a minimum, provide a framework for being more explicit about such extrapolation.

Future accountability assessments must rely on data collection that focuses on both the intervention and control groups. Outcomes (e.g., pollution and health measures) will need to be available for both populations, and data on relevant confounders is imperative to ensure that control areas can be adjusted to represent intervention areas.

### 2.3 Other Issues

The preceding sections outline the most salient issues for conducting causal accountability assessment: explicit definition of the research question, identification of an appropriate control population, and confounding. Many other issues relevant to accountability assessment require novel causal-inference methods. For example, the likelihood that interventions targeting specific pollutants also impact other pollutants presents the need for multipollutant statistical methods, which have begun to be considered in the accountability context [Dominici et al. 2010; Zigler et al. 2012]. For example, the causal effect of the PM\textsubscript{10} nonattainment designations on ambient O\textsubscript{3} can be
defined analogously to the causal effect of PM$_{10}$ nonattainment designations on PM$_{10}$. Complex interactions or synergies related to how multiple pollutants impact health can be suggested by knowledge that the greatest causal effects on health are evident in areas that exhibit the greatest causal effects on multiple pollutants. For example, one motivation behind [Zigler et al.] (2012) is to determine the extent to which causal effects of PM$_{10}$ nonattainment designations on health are most pronounced in areas for which PM$_{10}$ and O$_3$ were causally reduced by the PM$_{10}$ nonattainment designations.

More generally, accountability studies may aim to characterize “causal pathways” that describe how an air-quality regulation impacts health. Different causal pathways could represent multipollutant pathways or different links in the chain of accountability [Health Effects Institute] (2003). Methods for generating evidence of causal pathways characterizing the extent to which an intervention “directly” impacts health outcomes or “indirectly” impacts health “through” reducing air pollution is currently an active area of methodological research [Robins and Greenland] (1992), [Frangakis and Rubin] (2002), [Rubin] (2004), [VanderWeele and Vansteelandt] (2009). The key difficulty for such analyses is that pollution is a posttreatment concomitant variable (i.e., a variable that “lies on the causal pathway” between interventions and health), and standard adjustment of such variables (e.g., regression adjustment) is known to distort estimation of causal effects [Rosenbaum] (1984).

Finally, another technical difficulty pertains to the notion of “interference” between areas because interventions that impact air quality in a particular area are likely to affect pollution and health in nearby areas. This is an especially recent area of methodological work [Rosenbaum] (2007), [Hudgens and Halloran] (2008), [Tchetgen and VanderWeele] (2012). Causal inference with interference is aided by knowledge of exactly how interventions are expected to impact other areas (e.g., how far in space will the effects extend) and by knowledge of whether interference is considered when “assigning” interventions (e.g., areas thought to contribute to NAAQS violations in other areas are designated as nonattainment).

### 3 Conclusion

Recent emphasis on accountability assessment presents a unique opportunity to expand the use of potential-outcomes methods that explicitly define causal effects as consequences of actions. Increased precision with regard to causality demands new methods that formalize accountability studies as approximate experiments, which generally requires different tools than those common to air pollution epidemiology. Recognizing the demands of conducting accountability assessment under this framework for causal inference can inform the assessment of past and future air quality interventions.

### Acknowledgements

Funding for this work was provided by the Health Effects Institute.

### References


Figure 1: Schematic representation of causal effects as comparisons between observed trends and what would have been observed under alternative intervention scenarios.

(a) Temporal Stability Assumption

(b) Long-Term Changes