Causal Inference Methods for Estimating Long Term Health Effects of Air Quality Regulations

Corwin Matthew Zigler[*] Chanmin Kim [*] Christine Choirat [†]

John Barrett Hansen[†] Yun Wang [*] Lauren Hund [‡] Jonathan Samet[§]

Gary King[†] Francesca Dominici[*]

[*]Harvard TH Chan School of Public Health
[†]Harvard University
[‡]University of New Mexico
[§]Keck School of Medicine, University of Southern California

Contents

Abstract **4**

Introduction **7**

Overview of this Report 8

Case Study 1: PM_{10} Nonattainment Designations 8

Case Study 2: Scrubber Installations on Coal-Fired Power Plants 9

Specific Aims **10**

Methods and Study Design **11**

Publicly-Available Data and Reproducible Research for Accountability Assessment 11

Ambient Monitoring Data 11

Medicare Health Outcomes Data 12

Regulatory Data: EPA Green Book and Power Plant Emissions Controls Data 12

Other Supporting Data Sources 13

Existing Data Linkage 14

The Harvard Dataverse 15

AREPA R Package 15

Statistical Perspectives for Causal Accountability Assessment 16

Direct vs. Indirect Accountability 17

Potential Outcomes Methods: Framing as a Hypothetical (approximate) Randomized Experiment . 19

Methods for Confounding Adjustment: Propensity Scores 21

"Causal Pathways" Analyses: Causal Mediation Analysis and Principal Stratification 22

Results **26**

Case Study 1: Accountability Assessment of PM_{10} Nonattainment Designations 26

Linked Data Sources . . . 27
Defining the Intervention for Direct Accountability Assessment: Initial PM_{10} Nonattainment Designations . . . 29
Potential Outcomes Approach and Causal Effects of Interest . . . 31
Estimation of Propensity Scores for Confounding Adjustment . . . 33
Checking Covariate Balance . . . 35
Models for Estimating Causal Effects . . . 38
Case Study 1 Results . . . 38
Unadjusted Comparisons . . . 39
Average Causal Effects on Average Annual Ambient PM_{10} during 1999-2001 . . . 39
Average Causal Effects on Medicare Health Outcomes . . . 41
Associative and Dissociative Effects . . . 41
Conclusion and Discussion of Case Study 1 . . . 44
Additional work in Progress . . . 48
Case Study 2: Accountability Assessment of Power Plant Emissions Controls . . . 48
Linked Data Sources . . . 50
Defining the Intervention: SO_2 Scrubber Installation . . . 51
Defining Potential Outcomes for Principal Stratification and Causal Mediation Analysis . . . 52
Estimation: New Methods for Bayesian Nonparametric Mediation Analysis . . . 56
Case Study 2 Results . . . 57
Conclusion and Discussion of Case Study 2 . . . 61
Additional Work in Progress . . . 65

Discussion and Conclusions **66**

Abstract

Introduction: The regulatory environment surrounding air pollution control policies warrants a new type of epidemiological evidence. Whereas air pollution epidemiology has typically informed policies with estimates of exposure-response relationships between pollution and health outcomes, these estimates alone cannot support current debates surrounding the actual health impacts of air quality regulations. Directly evaluating specific control strategies is distinct from estimating exposure-response, and increased emphasis on estimating the effectiveness of well-defined regulatory interventions will enhance the evidence supporting policy decisions. The goal of this report is to provide new analytic perspectives and statistical methods for what we refer to as *direct accountability assessment* of the effectiveness of specific air quality regulatory interventions. Towards this end, we sharpen many of the distinctions surrounding accountability assessment initially raised in HEI Accountability Working Group (2003) through discussion, development, and deployment of statistical methods for drawing causal inferences from observational data. The methods and analyses presented here are unified in their focus on anchoring accountability assessment to estimating the causal consequences of well-defined actions or interventions. We discuss these analytic perspectives in the context of two direct accountability case studies considering different links of the so-called "Chain of Accountability" (HEI Accountability Working Group, 2003).

Methods: The statistical methods described herein consist of both established methods for drawing causal inference with observational data and of newly-developed methodology for causal accountability assessment. We sharpen the analytic distinctions between studies that directly evaluate the effectiveness of specific policies and those that estimate the exposure-response relationship between pollution and health. We emphasize how a potential-outcomes paradigm for causal inference can elevate current policy debates with more direct evidence of the extent to which complex regulatory interventions impact pollution and health outcomes. We outline the potential outcomes perspective and promote its use as a means to frame observational studies as approximate randomized experiments. Newly-developed methods for causal accountability assessment draw upon propensity scores, principal stratification, causal mediation analysis, spatial hierarchical models, and Bayesian estimation. The first of two illustrative case studies uses health outcomes among approximately 4 million Medicare benefi-

ciaries to estimate the causal health impacts of areas being designated as in nonattainment with the 1987 National Ambient Air Quality Standards for PM_{10} . The second case study quantifies the extent to which SO_2 scrubbers on coal-fired power plants causally affect emissions of SO_2 , NO_x , and CO_2 , as well as the extent to which causal emissions reductions mediate the causal effect of a scrubber on ambient concentrations of $PM_{2.5}$. Both case studies are anchored to our compilation of national, linked data on ambient air quality monitoring, weather, population demographics, Medicare hospitalization and mortality outcomes, continuous emissions monitoring for electricity-generating units, and a variety of regulatory control interventions. The resulting database has unprecedented accuracy and granularity for conducting the types of accountability assessment contained in this report, and a key component of our methods development for accountability assessment is the creation of tools for distribution of our linked data base and for reproducible research.

Results: The presentation of the PM_{10} nonattainment case study focuses on illustrating the most fundamental features of a causal-inference perspective on direct accountability assessment. Results of this case study indicate that, among areas designated as nonattainment during 1991-1995 the nonattainment designations causally reduced all-cause Medicare mortality and respiratory-related hospitalizations relative to what would have occurred absent the designations. The power plant emissions case study develops and illustrates our newly-developed statistical methods for multipollutant accountability assessment designed to quantify the causal pathways through which a regulatory action impacts ambient air quality. Results of the power plant case study indicate that presence of an SO_2 scrubber causally reduces ambient $PM_{2.5}$, that causal reductions in SO_2 emissions mediate approximately one-third of this effect, and that nearly two-thirds of the causal $PM_{2.5}$ reduction can be attributed to factors other than causal effects on SO_2 emissions.

Conclusion: Grounding accountability research in a potential outcomes framework and applying our new methods to our collection of national data sets, provides additional sound evidence of the health effects of long-term and large-scale air quality regulations, adding to that previously reported. Augmenting the existing body of research with rigorous evidence of the causal effects of well-defined actions will ensure that the highest-level

epidemiological evidence continues to support regulatory policies. Ultimately, our research provides support to the EPA and other stakeholders for incorporating health outcomes research into policy development.

Introduction

The claim that exposure to ambient air pollution is harmful to human health is hardly controversial in this day and age, due in large part to the evidence amassed through decades of air pollution epidemiological research. This body of research historically focused on hazard identification and more recently estimation of exposure-response (or, more formally, concentration-response) functions relating how health outcomes differ with spatial and/or temporal variations in ambient pollution exposure (Dockery et al., 1993; Pope III, 1996; Friedman et al., 2001; Krewski et al., 2003; Laden et al., 2006; Zeger et al., 2008; Pope III et al., 2009; Correia et al., 2013; Chen et al., 2013). Although considerable uncertainty remains with regard to essential finer-grade issues such as the specific shape of the exposure-response functions, the mechanics of exactly *how* pollution harms the human body, and the achievement of an "adequate margin of safety" dictated by the US Clean Air Act (CAA), evidence of the exposure-response relationship between pollution and health has motivated a vast array of air quality control policies in the US and abroad. The collection of these measures has undeniably improved ambient air quality over the past several decades (U.S. EPA, 2009; Samet, 2011).

Despite the success of such regulatory policies for cleaning the air, an evolving regulatory and political environment is placing new demands on input from the scientific community. With the prospect of increasing costs resulting from proposed tightening of air quality standards, the evidence motivating these policies is being subject to unprecedented scrutiny, and the scientific community must adapt by providing new types of evidence to support current and future regulatory strategies (Samet, 2011; Dominici et al., 2014; Zigler and Dominici, 2014). Policy makers, legislators, industry, and the public increasingly emphasize questions of whether past efforts have actually yielded demonstrable improvements to public health, whether the costs associated with implementation of control policies such as the Clean Air Act (e.g., annual costs of the 1990 Amendments reaching $65 billion by 2020 (U.S. EPA, 2010)) are justified, and which existing strategies have provided the greatest health benefits. These considerations reflect a shifting demand towards evidence of *effectiveness* of specific regulatory interventions. Starting most notably with a 2003 report from the Health Effects Institute (HEI Accountability Working Group, 2003), questions of so-called *accountability assessment* - assessment of the extent to which regulatory actions taken to

control air quality impact health outcomes - have been propelled to the forefront of policy debates. A National Research Council report commissioned by the US Congress recommended that an enhanced air quality management system strive to take a more performance-oriented approach by tracking effectiveness of specific control policies and creating accountability for results, with similar calls for the importance of accountability echoed by others, including EPA (National Research Council, 2004; Hidy et al., 2011; Hubbell, 2012; U.S. EPA, 2013). Increased emphasis on the direct study of the effectiveness of specific actions is one essential avenue to ensuring that epidemiological research continues to inform air quality control policies amid the current regulatory climate.

Overview of this Report

The goal of this report is to provide new analytic perspectives and statistical methods for what we refer to as *direct accountability assessment* of the effectiveness of specific air quality regulatory interventions. Towards this end, we sharpen many of the distinctions surrounding accountability assessment initially raised in HEI Accountability Working Group (2003) through discussion, development, and deployment of statistical methods for drawing causal inferences from observational data. The methods and analyses presented here are unified in their focus on anchoring accountability assessment to estimating the causal consequences of well-defined actions or interventions. We discuss these analytic perspectives in the context of two direct accountability case studies considering different links of the so-called "Chain of Accountability" (HEI Accountability Working Group, 2003). The statistical methods described herein consist of both established methods for drawing causal inference with observational data and of newly-developed methodology for causal accountability assessment.

Case Study 1: PM_{10} Nonattainment Designations

As a result of the 1990 amendments to the CAA, EPA began officially designating US counties as nonattainment for PM_{10} if 1) at least one pollution monitor in the county indicated a violation of the 1987 NAAQS for PM_{10} during the three previous years or 2) part of the county was thought to contribute to a violation of the NAAQS for PM_{10} in another area during that period. A county nonattainment designation induced the state containing that

county to submit a State Implementation Plan (SIP) detailing a strategy to achieve the NAAQS by a target date, and all counties not designated as nonattainment are considered in attainment and not required to produce a SIP. The Case Study 1: Accountability Assessment of PM_{10} Nonattainment Designations Section of this report presents an analysis of the extent to which initial PM_{10} nonattainment designations causally impacted ambient PM_{10} and health outcomes among Medicare beneficiaries. The presentation of this case study focuses on illustrating the most fundamental features of a causal-inference perspective on direct accountability assessment. The analysis focuses on three links in the Chain of Accountability: regulatory action, ambient air quality, and human health response.

Case Study 2: Scrubber Installations on Coal-Fired Power Plants

The 1990 amendments to Title IV of the CAA established the Acid Rain Program (ARP), which instituted a requirement for major emission reductions of both SO_2 and NO_x from stationary pollution sources. One goal of this program was to reduce total SO_2 emissions by ten million tons relative to 1980 levels (29.5 million tons per year). This drop was to be achieved mostly through cutting emissions from electricity-generating units (EGUs). One integral strategy to achieve the emissions-reduction goals of the program, especially among coal-fired EGUs, was the installation of flue-gas desulfurization technology ("scrubbers") for reducing emissions of SO_2 . The Case Study 2: Accountability Assessment of Power Plant Emissions Controls Section of this report presents an analysis of the extent to which SO_2 scrubbers on coal-fired power plants causally affected emissions of SO_2 , NO_x , and CO_2 , as well as ambient concentrations of $PM_{2.5}$. The focus of this case study is the illustration of our newly-developed statistical methods for multipollutant accountability assessment designed to quantify the causal pathways through which a regulatory action impacts ambient air quality. This analysis focuses on three links of the Chain of Accountability: regulatory action, emissions, and ambient air quality.

Specific Aims

A.1: Use a potential outcomes framework to define causal effects of interest for single pollutant accountability assessment and develop methods for estimation (year 1). We will develop a causal inference method for accountability research that uses principal stratification (Frangakis and Rubin, 2002) to isolate the causal pathways leading from regulation to changes in air quality and health. The proposed method will allow us to quantify and disentangle causal effects of the regulation on health that are 1) associated with causal effects of the regulation on air quality and 2) associated with causal pathways capturing other factors that do not involve changes in air quality (see Figure 3).

A.2: Define causal effects for multipollutant accountability assessment and develop methods for estimation (years 1-2). Current statistical methods for assessing the consequences of air quality management rely on specification of a single pollutant and on estimation of the health effects of that pollutant. We propose a method for multipollutant accountability research to estimate the joint effect of a regulation on multiple pollutants, allowing estimation of the (possibly synergistic) downstream effects on health.

A.3: Develop national databases, conduct epidemiological studies, and disseminate software and results (all years). We will assemble and link national data sets that will provide information on regulatory actions, ambient levels of criteria pollutants, health outcomes, and confounders for the entire U.S.. We will apply our proposed methods to our national data sets to estimate the impact on health indicators of separate regulations that target different pollutants. The necessary software and computational tools for our methods will be disseminated in conjuncture with results from our epidemiological studies.

Methods and Study Design

Publicly-Available Data and Reproducible Research for Accountability Assessment

A key component of our development of methods for accountability assessment has been the creation of data sources and tools for reproducible research. We have created a national and linked database containing information on ambient air quality monitoring, weather, population demographics, Medicare hospitalization and mortality outcomes, EPA nonattainment designations, continuous emissions monitoring for over 4,000 power-generating units in the US, and emissions control technologies employed at these units. Information in the data base spans the years 1990 - 2015 and exhibits unprecedented accuracy and granularity for conducting the types of accountability assessment contained in this report.

Our efforts towards transparency and reproducibility fall in three areas. First, R packages to implement the newly developed methods are currently in process, and will be made available on the Comprehensive R Archive Network (CRAN). Second, separate R packages are in development for the purposes of downloading, pre-processing, and linking the data sources described below. With the exception of Medicare data, all data sources used in this report are freely available and downloadable. The Medicare data are publicly available, but must be purchased. The programs we are currently developing will allow anyone with R to automatically download and integrate the freely-available data sources for use in their own research. Finally, we are working to make the specific data sets used for our analyses (with appropriate privacy protections for health data) accessible through the Harvard Dataverse, an online repository for sharing, citing, and preserving research data (King, 2007; Crosas, 2011). Distribution of our database and all software programs will permit any given study based on our database and software to be replicated using the appropriate database and software versions.

Ambient Monitoring Data

Air Quality System (AQS) from EPA (`http://www.epa.gov/ttn/airs/airsaqs/`). We have developed scripts that retrieve daily and annual data at the monitor level. These data, pre-processed by the EPA

from hourly raw data, contains measurements corresponding to over 10,000 monitors for the period 1990 to 2014, approximately 5,000 of which are currently active. Specifically, our scripts allow for obtaining data on: criteria gases (O_3, SO_2, CO, and NO_2); particulate matter ($PM_{2.5}$, $PM_{2.5}$ non-FRM, and PM_{10} Mass, and $PM_{2.5}$ speciation) meteorological factors (wind speed, temperature, barometric pressure, dew point); and toxics (HAPS and VOCs) and lead.

Medicare Health Outcomes Data

Centers for Medicare and Medicaid Services (CMS) (`http://cms.hhs.gov/`). Data are available for 1999-2015. These data provide restricted-access daily data at the zip code level.

- *Cohort*: All Medicare enrollees by year of enrollment, including age, gender, race, state, 5- and 9-digit zip code identifiers for their residence (40,000,000 people per year).
- *Mortality*: Date of death for enrollees within the Medicare cohort.
- *Hospitalizations*: Hospitalization records for all Medicare enrollees, including date of hospitalization, length of stay in hospital, International Classification of Diseases (ICD) primary and secondary diagnostic and procedure codes associated with the hospitalization, and the costs billed to Medicare for the hospitalization (see Dominici et al. (2006); Peng et al. (2008); Zanobetti et al. (2014)) for details).

Regulatory Data: EPA Green Book and Power Plant Emissions Controls Data

County-level nonattainment designations are available from the EPA Green Book for all criteria pollutants since 1978 (`http://www.epa.gov/airquality/greenbook/data_download.html`). For the period 1995-2012, open-access daily data are available at the EGU and power plant level from EPA's Air Markets Program Data (AMPD) (`http://ampd.epa.gov/ampd/`), where a power plant is defined as a facility with one or more EGUs. We have collected data on 4,164 units belonging to 1,248 facilities, the totality of facilities that participated in the Acid Rain Program. As shown in Figure 1, these monitored facilities are an important source of SO_2 emissions: they

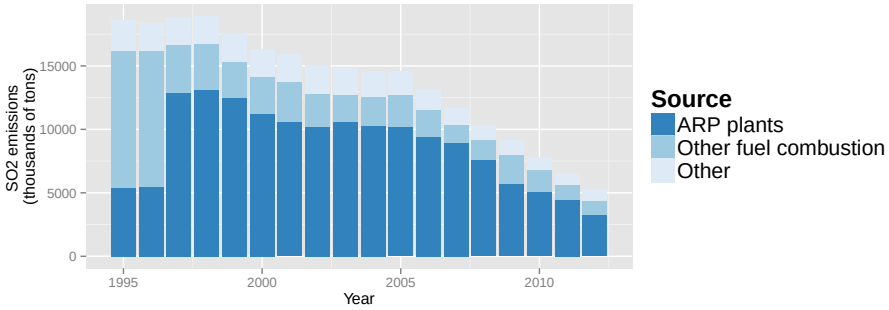


Figure 1: Relative importance of SO_2 emissions from ARP-monitored facilities.

represent approximately 20% of all US power plants, but account for 75% of fuel combustion emissions and 65% of overall emissions.

Specifically, we have collected information about each unit (e.g., state, county, latitude, longitude); ARP phase (I, II, opt-in, substitution, compensating); SO_2, NO_x, and CO_2 emissions; average emission NOx rate; heat input, gross load, steam load, operating time, and status; primary and secondary fuel types (e.g., coal, diesel oil, natural gas); and scrubber technologies (whether a scrubber is installed and, if so, the technology it uses) for SO_2, NO_x and particulate matter (PM).

Other Supporting Data Sources

We obtain population demographic information for the year 2000 from the **US Census Bureau** (`http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml`). We obtain annual temperature data from the **National Climactic Data Center** (`http://www.ncdc.noaa.gov/cdo-web/`). County-level smoking rates are obtained from the CDC Behavioral Risk Factor Surveillance System (`http://www.cdc.gov/brfss/data_tools.htm`). From the **US Energy Information Adiministration (EIA)** (`http://www.eia.gov/`) we have retrieved monthly and annual data (from forms EIA-767, EIA-906A and EIA-923) for the period 1985-2012 at the power plant level, including: the year that each unit was or is expected to be in compliance; strategy for compliance for PM; actual or projected in-service and retirement dates; primary fuels and alternate fuel capacity; monthly net electrical generation; fuel; and monthly heat, ash, and sulfur contents. We have included EIA information in the AMDP database, relying on the unique facility and unit identifier that unambiguously corresponds to a power-generating

unit.

Existing Data Linkage

Linking data sources and dealing with spatial misalignment. The data sources that we are using are reported for different geographical units. ARP and AQS data are coordinate-based, whereas Medicare enrollee data is available at the zip code level. Data linkage strategies have to take into account spatial misalignment.

AQS and Medicare. In order to link ambient monitoring sites to zip codes, which is the basis for linking AQS data to Medicare health outcomes, we consider each zip code in the U.S. and enumerate all EPA monitoring sites within a certain distance. When a zip code is close to more than one monitor, that zip code is linked only to the closest monitor. After each zip code is linked to at most 1 monitoring site, Medicare data is then obtained at the zip code and aggregated to level of the monitoring site by combining data on all zip codes assigned to each monitor. Our process ensures that each zip code is reliably assigned to only one monitoring site, is computationally efficient, customizable (e.g., could be used to link zip codes at various distances) and decouples the zip code-AQS linkage from the combination with Medicare data so that the linkage can be conducted without any specialized knowledge of or access to Medicare data.

AMPD and Medicare and AMPD and AQS. Our linkage algorithm can also be used to link AMDP power plants to AQS monitoring sites or zip codes for our analyses of the Acid Rain Program. Our goal is to provide computer programs and source code for the monitor-zip code linkage that could be re-implemented for any specified set of monitoring locations (e.g., all PM_{10} monitoring sites with "population oriented" monitors, or all monitoring sites containing monitors for both PM_{10} and O_3). Another data process improvement pertains to inclusion/exclusion criteria for monitoring sites, for example we are able to reliably exclude monitors that are calibrated on a micro-scale, or those for which there are not enough daily measurements for a given year.

The Harvard Dataverse

Our datasets and databases (linking EPA and EIA sources and simulated Medicare information) are accessible from a restricted-access Harvard Dataverse (`https://dataverse.harvard.edu/airqualregs`), an online repository for sharing, citing, and preserving research data (see King (2007)). Upon completion of our database compilation and analyses, access to the Dataverse will become unrestricted.

AREPA R Package

We are also developing software programs within the R statistical environment to allow for easier and broader distribution of our results. In particular, we are developing the `arepa` (EPA data retrieving and processing) package in a private GitHub repository (`https://github.com/czigler/arepa`). Our codebase leverages on the wealth of tools provided by R, more specifically: (a) fast and efficient in-memory big data manipulation with `data.table` `http://cran.r-project.org/web/packages/data.table/index.html`); (b) GIS capacities with `sp` (`http://cran.r-project.org/web/packages/sp/index.html`). The `arepa` repository will become public and the package will be made available on the Comprehensive R Archive Network CRAN `http://cran.r-project.org/` alongside completion of the papers.

The `arepa` package is currently used within our Group and provides three main groups of functionalities to improve the efficiency and reproducibility of our workflow.

1. Script-based downloads for daily and annual AQS data, as described in Ambient Monitoring Data Section.

2. Spatial linkage procedures that implement the methods of the Existing Data Linkage Section. For example, figure 2 illustrates the spatial linkage between AMPD power plants and AQS $PM_{2.5}$ monitors in 2010 with a default radius of 100 km. Blue circles correspond to a 100 km radius around a power plant that has been successfully linked to a unique monitor, while grey circle corresponding to power plants that have been discarded for lack a unique monitor within the 100 km range.

3. Creation of an indexed dataset from which Medicare data can be retrieved at the zip code level and then

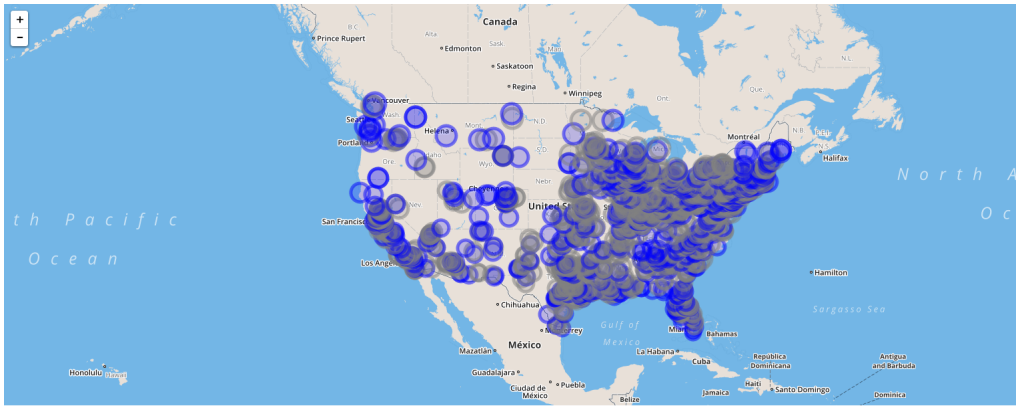


Figure 2: Spatial linkage between AMPD power plants and AQS $PM_{2.5}$ monitors in 2010

aggregated around AQS monitors or AMPD power plants. We will provide simulated Medicare data to illustrate the data formats used in our routines so that they can be used by other research groups.

Statistical Perspectives for Causal Accountability Assessment

The role of causality is of obvious import for informing policy decisions, and the causal validity (or lack thereof) of epidemiological evidence has always been central to the integration of scientific evidence into policy recommendations (U.S. EPA, 2009). However, approaches to inferring causality from available observational data can vary depending on the scientific question of interest and the data available for analysis.

Causal inference in air pollution epidemiology has most commonly been undertaken within a "classical" paradigm, which construes causal validity on a continuum according to how likely an observed association (e.g., between pollution and health) can be interpreted as "causal" (Glass et al., 2013). This continuum is explicitly considered in the approach to Integrated Science Assessments conducted by EPA, which classify evidence of the association between pollution exposure and health as a "causal relationship," "likely to be a causal relationship," "suggestive of a causal relationship," "inadequate to infer a causal relationship," or "not likely to be a causal relationship." Even in the absence of the word "causal," the bulk of air pollution epidemiology has been implicitly undertaken with this classical approach; an exposure-response relationship between pollution and health is estimated (e.g., in a cohort study), then a judgment is made as to whether this relationship can be reasonably

interpreted as causal, and finally, hypothetical changes in exposure are input into the exposure-response function to infer the resulting "health effect" that would be caused by such a change in pollution. Many such studies have been integral to issues of accountability (Dockery et al., 1993; Laden et al., 2006; Zeger et al., 2008; Pope III et al., 2009; Correia et al., 2013).

As an alternative to the classical paradigm, the potential-outcomes paradigm for causal inference has the distinctive feature that causal effects are explicitly defined as consequences of specific actions (Rubin, 1978). Rather than infer causality based on belief of whether an estimated exposure-response relationship can be interpreted as causal, potential-outcomes methods entail definition of a clearly-defined action (a "cause"), the effects of which are of interest. Some existing accountability assessments have been (often implicitly) undertaken within a potential-outcomes paradigm for causal inference, the common thread being application of the core tenets of experimentation to observational settings (Pope III et al., 2007; Moore et al., 2010; Currie and Walker, 2011; Rich et al., 2012; Chen et al., 2013; Friedman et al., 2001; Hedley et al., 2002; Clancy et al., 2002; Tonne et al., 2008; Chay et al., 2003; Greenstone, 2004; Zigler et al., 2012; Deschenes et al., 2012).

Direct vs. Indirect Accountability

Studies framed as accountability studies can be classified according to the specific scientific question of interest. Studies that answer questions of the form: "What is the relationship between exposure to pollution and health outcomes?" can be aptly described as *indirect accountability* studies (Zigler and Dominici, 2014). This type of question has been at the center of air pollution epidemiology for decades, and answers typically come in the form of exposure-response relationships between (changes in) pollution exposure and (changes in) health outcomes. Importantly, these studies do not consider the effectiveness of any specific regulatory action, but provide valuable evidence for indirectly predicting the impact of policies. For example, EPA routinely uses exposure-response estimates to estimate the expected benefits of current and future policies; if a policy reduces (or is expected to reduce) pollution by a certain amount, then the exposure-response relationship indirectly implies the health impact of the policy insofar as the relationship can be deemed causal (U.S. EPA, 2009, 2010, 2012). Indirect accountability assessments typically assume that any observed exposure-response relationship would

persist amid the complex realities of actual regulatory implementation that will typically impact a variety of factors. As a consequence, health impacts of regulatory interventions may not be accurately characterized by indirectly applying exposure-response estimates to accountability assessments.

We focus on a slightly different perspective on accountability assessment that we describe as *direct accountability* (Zigler and Dominici, 2014). Direct accountability studies target a different scientific question than studies of exposure-response relationships. Rather than investigate the relationship between pollution and health, these studies answer the question "What is the relationship between a specific regulatory intervention and health?". These studies are "direct" accountability studies in that they directly evaluate the effectiveness of well-defined regulatory actions, which more definitively informs questions as to the actual health benefits of these actions. While relatively less common to air pollution epidemiology than studies of exposure-response, we argue that direct accountability assessments are best equipped to meet the demands of a shifting regulatory environment wrought with questions surrounding the effectiveness of specific policies. Of particular importance is the noted lack of direct evaluations of broad, complex regulatory interventions, which are of utmost relevance to policy debates (Health Effects Institute, 2010; van Erp et al., 2012).

The analytic perspectives and statistical methods described in this report, namely, those rooted in potential-outcomes methods for causal inference, are particularly well suited to answering questions of direct accountability. The purpose for distinguishing between direct and indirect accountability is not to highlight the need for "causal" versus "associational" evidence, as all research to provide such evidence shares the goal of establishing causality. Rather, we argue that the shifting regulatory environment would be better informed by evidence of the effectiveness of specific control policies, and that traditional epidemiological approaches tailored to exposure-response estimation are not the most direct means to provide this evidence. In an environment that brings skepticism and doubt about results drawn from observational data, analyzing specific interventions with approaches rooted in potential-outcomes thinking can clarify the basis for drawing causal inferences and bring a higher level of credibility to evidence used to support policy decisions (Dominici et al., 2014; Zigler and Dominici, 2014).

Potential Outcomes Methods: Framing as a Hypothetical (approximate) Randomized Experiment

The underlying features of randomized studies that make them the "gold standard" for generating causal evidence remain pertinent to causal accountability assessment, with potential-outcomes methods framing observational studies according to how well they can approximate randomized experiments (Rubin, 2008; Hernan et al., 2008). The key idea is to define a (possibly hypothetical) experiment consisting of an "intervention condition" and a "control condition" such that if populations could be randomly assigned to these conditions, differences in observed health outcomes would be interpreted as causal effects of the intervention. While defining the intervention condition in accountability studies can be straightforward (e.g., it will likely be a regulatory action that actually occurred), framing accountability as a hypothetical experiment forces the specification of some alternative action that might have otherwise occurred to serve as a relevant control condition. This exercise formalizes the research question by explicitly defining a causal effect as a comparison between what would happen under well-defined competing conditions. Hence the name of the potential outcomes paradigm; a causal effect of "Action A" relative to "Action B" is defined as the comparison of the *potential outcome* if "Action A" were taken with the *potential outcome* if "Action B" were taken. Thus, the salient question for accountability is not "Did health outcomes change after the intervention?" but rather "Are health outcomes different after the intervention than they would have been under a specific alternative action?". For example, the case study in the Case Study 1: Accountability Assessment of PM_{10} Nonattainment Designations Section attempts to answer the question: "Are Medicare health outcomes in PM_{10} nonattainment areas different than they would have been if the nonattainment designations had never occurred?" The hypothetical treatment condition is the set of observed PM_{10} nonattainment designations, while the hypothetical control condition is the setting where these areas were never designated.

Of utmost importance is that definition of the causal effect of interest is purely conceptual and conducted *without regard to any assumed statistical model*. Different models could be used to actually *estimate* this effect, but the effect itself, along with its interpretation, remains consistent regardless of the modeling approach. This clarity is essential for producing policy-relevant evidence. Compare this to traditional epidemiological studies, which a) do not necessarily explicate an action defining effects of interest and b) define "health effects" with

parameters (e.g., regression coefficients) in a statistical model.

Estimating causal effects with comparisons between potential outcomes under competing intervention and control conditions is met with the fundamental problem that if the intervention is enacted, then outcomes under the control condition are unobserved. For example, evaluating the effect of a PM_{10} nonattainment designations requires knowledge of what would have potentially happened if the the nonattainment designations had never occurred. Hypothetical scenarios that never actually occurred are often referred to as "counterfactual" scenarios, and estimating what would have happened under such scenarios is perhaps the most important challenge for direct accountability assessment.

Counterfactual scenarios have been explicitly considered, for example, in EPA cost-benefit analyses of the CAA mandated by Section 812 of the act, which project two counterfactual pollution scenarios: one that assumes past exposure patterns would have continued without the 1990 CAA Amendments and another that assumes an expected change in exposure patterns under full implementation of the 1990 Amendments. These projections are coupled with exposure-response functions from the epidemiological literature to project counterfactual health scenarios that form the basis of the health-benefits analyses (U.S. EPA, 2010, 2012). However, these counterfactual projections are not validated against studies of actual interventions, and thus are not sufficient for fully characterizing the relationships between regulatory strategies and health (HEI Accountability Working Group, 2003).

Rather than project counterfactual scenarios by combining assumed exposure patterns with exposure-response estimates, potential-outcomes approaches typically use actual data from the "control group" of the hypothetical experiment to learn about what would have happened under the hypothetical control condition, rendering identification of a control population of vital importance. When assessing the impact of regulatory intervention relative to what would have happened absent the intervention, control populations could be defined based on time (e.g., a population before promulgation of a regulation), or space (e.g., if some areas are subject to an intervention and others not). Whether outcomes in the control population can actually characterize what would have occurred absent the intervention boils down to the familiar concept of confounding, although what constitutes a confounder of the effect of an intervention is slightly different from the common conception of a confounder as something

that is related to both pollution exposure and health.

For direct accountability, a comparison between outcomes among the intervention and control conditions is unconfounded if the two populations are comparable with regard to factors that relate to outcomes. An unconfounded comparison of outcomes between the intervention and control conditions yields an estimate of the causal effect. If the two populations differ on important factors related to outcomes, then such a comparison is a convolution of differences due to the intervention and differences due to other factors. Thus, if an important factor relating to health, for example, smoking behavior, is comparable across the intervention and control populations, then smoking behavior is not a confounder in the assessment of the intervention. Compare this to the typical setting of exposure-response studies, where a "confounder" is generally regarded as a factor that is simultaneously associated with pollution exposure and health outcomes. In both settings, the definition of a confounder is a factor that is associated with "exposure" and "outcome," the key difference being that, in a direct accountability study, the "exposure" is actually the intervention.

Methods for Confounding Adjustment: Propensity Scores

There are a variety of analytic tools available to address confounding in nonrandomized accountability studies. Specialized study designs, often described as "quasi experiments" circumvent the need to consider confounding directly, as they support assumptions that an intervention was "quasi" randomized in the sense that it is unrelated to health outcomes (Greenstone and Gayer, 2009; Dominici et al., 2014). Absent the availability of such specialized circumstances, methods for confounding adjustment (e.g., matching, weighting, stratification, or standardization) adjust for differences between intervention and control populations so that comparison groups can be regarded as similar on the basis of observed factors, thus mimicking the design of a randomized study.

One broad class of methods for confounding adjustment relies on the *propensity score* (Rosenbaum and Rubin, 1983; Robins et al., 2000; Rubin, 2008; Stuart, 2010). Propensity score methods share the same objective as, say, adjusting for covariates in a regression model, but have been shown to have several benefits relative to reliance on parametric regression models. Propensity scores represent a dimension-reduction procedure where multivariate confounding information is condensed into a one-number confounder summary called the estimated propensity

score. Observations with similar values of the estimated propensity score can be regarded as similar on the basis of all of the covariates that were used for its estimation. The value of the estimated propensity score can then be used to adjust for confounding via matching, weighting, or sub classifying the observed sample based on the estimated propensity score to ensure groups of treated and control observations that are comparable. If the propensity score model is adequately specified (i.e., if there is no unmeasured confounding), then average outcomes in control observations represent what would have happened (on average) in treatment observations with similar values of the propensity score, just as would be the case in a randomized study. Our analysis in the Case Study 1: Accountability Assessment of PM_{10} Nonattainment Designations Section makes use of propensity scores for confounding adjustment, and we discuss additional considerations in the context of the analysis.

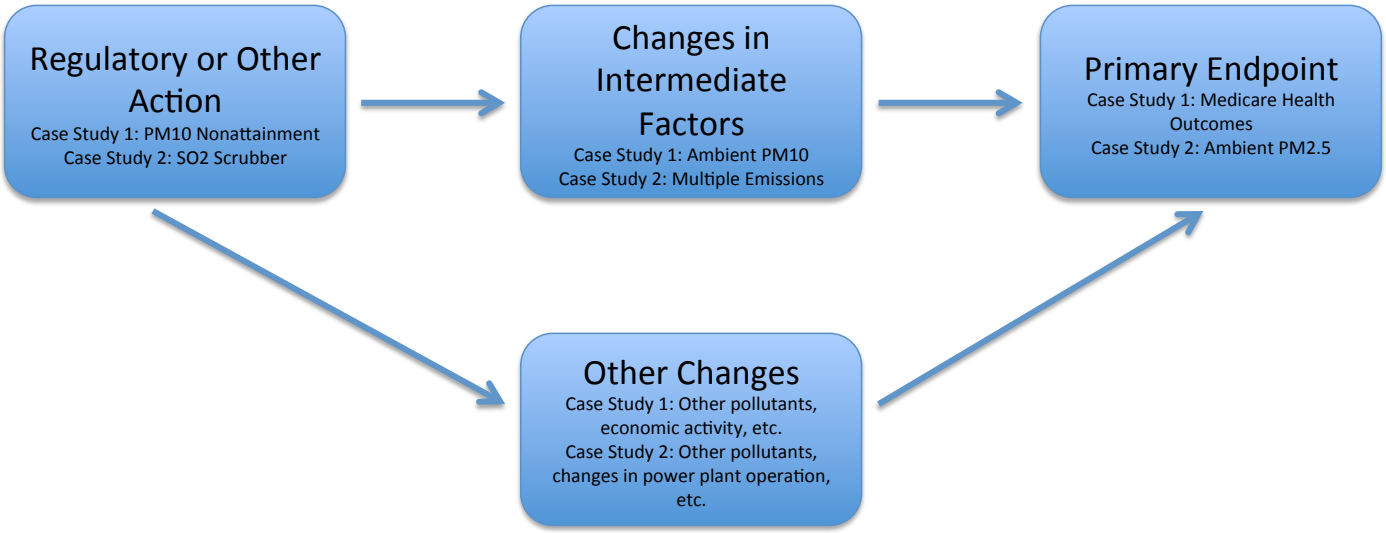
"Causal Pathways" Analyses: Causal Mediation Analysis and Principal Stratification

Another objective for causal accountability assessment that is in concert with the Chain of Accountability is to quantify the relative importance of the possible causal pathways that constitute links in the chain. For example, one set of questions may relate to the extent to which the causal effect of an intervention on health outcomes acts "through" reducing ambient pollution, or there may be questions regarding the extent to which an intervention effect on ambient pollution is mediated through specific emissions. Figure 3 presents a schematic representation of different causal pathways for accountability assessment.

Understanding the pathways through which an intervention affects ambient air quality or health outcomes is critical for informing policy decisions. From this perspective, intermediate factors in the Chain of Accountability that lie between the regulatory action and human health response can be regarded as lying "on the causal pathway." Because such intermediate outcomes are *posttreatment concomitant variables* that are expected to simultaneously be affected by the intervention and have bearing on outcomes, standard regression adjustments will not permit estimation of causal effects (Rosenbaum, 1984). We consider two related causal frameworks for characterizing causal effects with intermediate variables: causal mediation analysis and principal stratification.

Causal mediation analysis is a framework designed to isolate specific causal pathways in order to assess whether the causal effect of an intervention on an outcome is mediated through the causal effect of the intervention

Figure 3: Schematic description of direct and indirect causal pathways for accountability assessment. Air quality interventions are typically intended to impact primary pollution and health outcomes through reducing specific emissions and/or ambient pollutants (indirect effects) but can, in reality, impact outcomes through other causal pathways (direct effects).



on the intermediate variable (VanderWeele and Vansteelandt, 2009). Causal mediation analysis construes *two* hypothetical interventions in the context of accountability. First, is the hypothetical intervention representing the "treatment" and "control" groups described above. Second, causal mediation analysis also requires definition of an *additional* intervention that acts directly on the intermediate variable, independently of the first intervention. In the example of Case Study 2, the first hypothetical intervention represents the presence of an SO_2 scrubber on a coal-fired power plant, which is the actual regulatory intervention of interest. The second intervention is purely hypothetical, and corresponds to a way in which power plant emissions could be manipulated independently of a scrubber. Definition of potential outcomes based on these two interventions permits decomposition of the total causal effect of an intervention into effects that are *direct* and *indirect* effects (Robins and Greenland, 1992). A direct effect in the power plant context corresponds to the causal effect of the scrubber that acts directly on ambient pollution in that it is attributable to causal pathways not involving emissions. An indirect effect in this context is the effect of a scrubber on ambient pollution that can be attributed to causal emissions reductions. Causal mediation analysis is explained an illustrated with more technical detail in the Case Study 2: Accountability

Assessment of Power Plant Emissions Controls Section.

Principal stratification is a related approach for causal inference with intermediate variables that shares similar objectives to those of causal mediation analysis, but construes only one hypothetical intervention (i.e., the regulatory intervention being assessed) (Frangakis and Rubin, 2002). Principal stratification in the accountability context aims to quantify the extent to which causal effects of an intervention on the primary outcome coincide with causal effects of the intervention on the intermediate variable (Zigler et al., 2012). For example, Case Study 1 investigates the extent to which a causal effect of PM_{10} nonattainment designation on Medicare health outcomes coincides with the causal effect of the designation on ambient pollution. Towards this end, principal stratification defines two types of causal effects: dissociative and associative causal effects. In the nonattainment example, *dissociative effects* quantify the extent to which the intervention causally affects health outcomes when the intervention does not causally affect pollution. Dissociative effects are similar to direct effects in the mediation analysis framework in that they represent causal health effects of an intervention that are indicative of causal pathways other than ambient pollution. *Associative effects* quantify the causal effect of the intervention on health when the intervention causally affects ambient pollution, and are similar to (but distinct from) indirect effects in the causal mediation analysis framework. An associative effect that is large relative to the dissociative effect indicates that the causal effect of the intervention on health outcomes is greater in areas where pollution was causally affected than in areas where there is little or no effect of the intervention on pollution. This would suggest the presence of a causal pathway whereby the intervention impacts health through changing pollution. Dissociative effects that are similar in magnitude to associative effects indicate that the health impact of the intervention is similar regardless of whether the intervention causally impacted ambient pollution, which suggests the presence of other causal pathways through which the intervention affects health without changing pollution.

The theoretical and technical differences between principal stratification and mediation analysis have been closely examined in the causal inference literature (e.g., VanderWeele (2008); Joffe and Greene (2009); Pearl (2011); Rubin (2004)) in the setting of a single mediating factor. The Case Study 1: Accountability Assessment of PM_{10} Nonattainment Designations Section illustrates the use of principal stratification in our analysis of PM_{10} nonattainment designations. The Case Study 2: Accountability Assessment of Power Plant Emissions Controls

Section provides technical details of new methods for both principal stratification and causal mediation analysis in the multipollutant context, and interprets both analyses in the context of the case study.

Results

Case Study 1: Accountability Assessment of PM_{10} Nonattainment Designations

Here we employ the analytical perspectives outlined in the Statistical Perspectives for Causal Accountability Assessment Section to provide the first direct accountability assessment of one integral regulatory strategy defined under the 1990 amendments to the CAA: the initial designation of areas as nonattainment with the 1987 National Ambient Air Quality Standards (NAAQS) for PM_{10} . In contrast to initial efforts to examine the impact of the CAA as a whole (HEI Accountability Working Group, 2003; U.S. EPA, 2010), we provide a focused analysis of the initial PM_{10} nonattainment designations for two important reasons. First, the process whereby the EPA sets NAAQS and initiates nonattainment designations is one integral tool for managing air quality under the CAA, and quantifying the effects of this specific decision process can prove valuable. Second, a focused characterization of the health impact of a specific set of regulatory decisions provides more targeted direct accountability of the effectiveness of a specific regulatory decision, which yields more targeted evidence for informing future policies than overall assessments of the CAA.

Our analysis differs from traditional epidemiological investigations of the long-term association between pollution exposure and health in that we adopt an analytical perspective designed specifically to estimate causal effects of a specific set of actions, the nonattainment designations, rather than characterize the exposure-response relationship between pollution and health in the time frame that saw myriad regulatory actions contributing to improved air quality. Specifically, we use a principled causal-inference framework to assess whether the initial PM_{10} nonattainment designations caused improvements in Medicare health outcomes. In accordance with the Chain of Accountability, we view changes in ambient PM_{10} as intermediates on the causal pathway between regulatory decisions and health outcomes, representing three key features of the chain: regulatory action, ambient air quality, and human health response (HEI Accountability Working Group, 2003). In addition to estimating overall impacts of nonattainment designations on ambient PM_{10} and Medicare health outcomes, our approach provides additional information as to the relative importance of different causal pathways through which regulatory decisions may impact health.

Linked Data Sources

We assembled a national, linked database using the tools described in the Publicly-Available Data and Reproducible Research for Accountability Assessment Section to conduct our investigation. The study population consists of US Medicare beneficiaries living within 6 miles of a PM_{10} monitoring location in 2001. The locations used for the analysis are those EPA monitoring stations located in the Western US that were in operation at any point between 1990 and 2001 (see Figure 4(a)). The Western region was chosen because virtually all initial nonattainment designations for PM_{10} occurred in this part of the country. From the EPA Green Book, we enumerated every county in the US designated as nonattainment for PM_{10} between 1991 and 1995. Annual average ambient PM_{10} concentrations from 1990-2001 were obtained from pollution monitor locations in the EPA AQS database. Annual average PM_{10} concentrations were regarded as missing if the percentage of valid measurements was less than 67%. Health data was assembled from the Center for Medicare and Medicaid Services (CMS) Medicare Part A hospital claims and enrollment data. From the CMS enrollment file, we enumerate all Medicare beneficiaries residing in a zip code within 6 miles of a pollution monitor during 2001. Beneficiaries living within 6 miles of multiple monitors are linked to the monitor closest to their zip code of residence. Data available on Medicare beneficiaries includes basic demographic information, mortality information, and hospitalization records. Hospital billing claims data were used to identify hospitalizations for cardiovascular (CVD)-related and respiratory-related illness. CVD-related hospitalizations were defined as those having ICD9-CM codes 390.xx to 495.xx. Respiratory-related hospitalizations were defined as those relating to COPD (ICD9-CM 490.xx to 492.xx) or respiratory tract infections (ICD9-CM 464.xx to 466.xx and 480.xx to 487.xx). Individual-level health data are aggregated to the level of of the monitoring location to yield average demographic information (average age, % female, etc.) and outcome rates (mortality rate, hospitalization rates) for all beneficiaries living within 6 miles of each monitoring location.

We augment the pollution-health linked data base with county-level information from the 2000 US census and from the Center for Disease Control and Prevention Behavioral Risk Factor Surveillance Survey (CDC-BRFSS). County-level information includes population demographics, urbanicity, and smoking rates. Additionally, county-

level values of annual average daily maximum temperature in 1990 were obtained by averaging across monitoring stations available from the National Climactic Data Center. Table 1 summarizes the linked information obtained in the data base. For our analysis, data are considered at the monitor level, that is, for each monitoring location, we have a specific location (latitude and longitude), measures of ambient pollution, demographic characteristics of the county containing the monitor, and aggregated health information on all Medicare beneficiaries residing within a 6 mile radius. The initial analysis data set contains the 547 monitoring locations depicted in Figure 4(a), with health data comprised of information on 3,971,610 Medicare beneficiaries. Among these 547 locations, 268 are located in nonattainment areas, corresponding to 2,349,691 Medicare beneficiaries. Note that monitoring locations with fewer than 20 Medicare beneficiaries residing within 6 miles in 2001 were excluded.

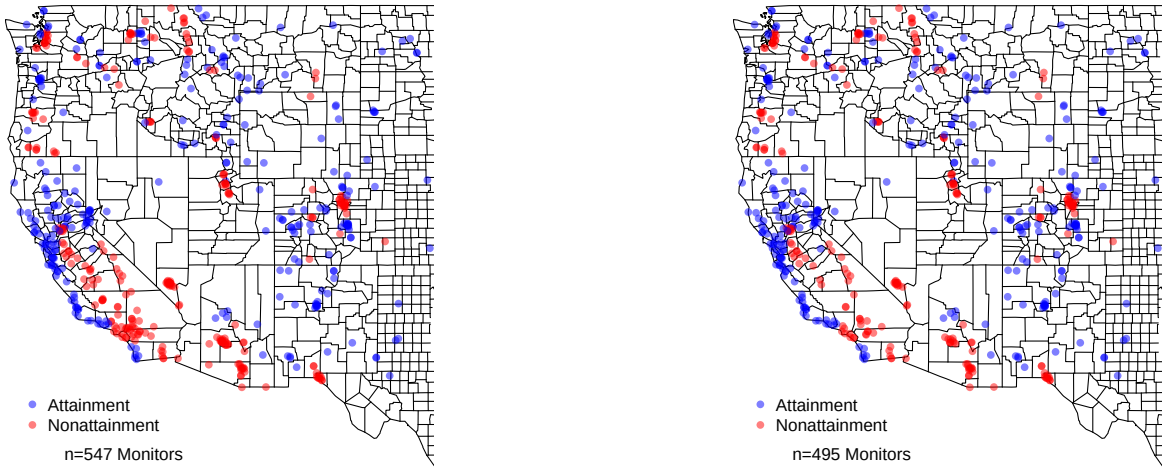
The outcome variables for our analysis are: 1) the average annual ambient concentration of PM_{10} during 1999-2001; 2) the all-cause mortality rate (# of deaths per beneficiary); 3) the CVD-related hospitalization rate (# of hospitalizations per person-year); and 4) the respiratory-related hospitalization rate (# of hospitalizations per person-year). Person-years are used in the analysis of the hospitalization outcomes to account for the fact that beneficiaries can die or un-enroll from Medicare during the year, and hospitalization records are only available during the time period of enrollment. In contrast, mortality is known regardless of Medicare enrollment status. All other variables listed in Table 1 are considered covariates in our analysis. Note that some covariate values are measured after the nonattainment designations: census variables are from the 2000 census, and Medicare demographic data are measured in 2001. We assume that such variables are not affected by the nonattainment designations and as such are reliable proxies for the same quantities in the years preceding the nonattainment designations.

Our analysis is confronted with two types of missing ambient pollution data. First, 284 locations (131 nonattainment) had missing PM_{10} measurements in 1990. Missing values of average annual ambient pollution in 1990 were singly imputed using posterior mean predictions from a spatial hierarchical random effects model as described in Appendix A.1. The second type of missing pollution data are missing average annual ambient concentrations during 1999-2001. Average ambient PM_{10} concentration for 1999-2001 was missing for 157 (70 nonattainment) locations. These follow up pollution measures are used as outcomes in our analysis, and are mul-

Figure 4: Locations of all 547 PM_{10} monitoring locations available for analysis and for the 495 locations retained after propensity score pruning

(a) Entire Monitor Set

(b) Pruned Monitor Set



tiply imputed from the models described in the Models for Estimating Causal Effects Section as a byproduct of our Bayesian estimation procedure. Missing values of county-level average annual daily maximum temperature were imputed for 61 (12 nonattainment) locations using k nearest neighbor mean imputation. All other covariates and outcomes listed in Table 1 were fully observed.

Defining the Intervention for Direct Accountability Assessment: Initial PM_{10} Nonattainment Designations

As a result of the 1990 amendments to the CAA, EPA began officially designating US counties as nonattainment for PM_{10} if 1) at least one pollution monitor in the county indicated a violation of the 1987 NAAQS for PM_{10} during the three previous years or 2) part of the county was thought to contribute to a violation of the NAAQS for PM_{10} in another area during that period. The first nonattainment designations occurred in 1991, and in the Western US considered in our case study, additional areas were designated as nonattainment through 1995 (see Figure 7). A

Table 1: Summary statistics for covariates and outcomes available for the analysis of PM_{10} nonattainment designations. Variables marked with * are those included in the model that estimates the propensity score and for additional covariate adjustment in models for pollution and health outcomes.

	Attainment Areas		Nonattainment Areas	
	Mean	SD	Mean	SD
Monitor Data				
Ambient PM10 1990*	26.36	6.55	39.2	12.75
Medicare Data				
Medicare Beneficiaries*	5813.33	9579.5	8767.5	13387.71
Age*	74.77	1.1	74.7	1.31
% Female*	0.55	0.05	0.56	0.06
% White*	0.9	0.14	0.87	0.14
% Black*	0.01	0.03	0.03	0.05
County-Level Data				
Population*	889937.31	1472176.42	3380578.87	5125807.84
Housing Density*	0.42	0.08	0.4	0.08
% Urban Living*	0.72	0.23	0.85	0.19
Median Income*	42148.87	10415.55	40873.47	7764.82
% High School Graduates*	0.84	0.06	0.8	0.09
5-Year Migration Rate*	0.23	0.06	0.21	0.06
Smoking Rate*	0.18	0.06	0.2	0.04
Annual Maximum Temperature*	65.75	6.6	72.39	10.14
% Hispanic	0.16	0.14	0.25	0.2
% White	0.73	0.18	0.64	0.21
% Black	0.02	0.03	0.04	0.03
% Female	0.5	0.01	0.5	0.01
Pollution and Health Outcomes				
Ambient PM10 1999-2001	21.58	6.43	31.56	13.28
Mortality Rate	62.58	16.95	62.51	12.4
Hospitalization Rate: CVD	83.74	24.24	92.09	26.65
Hospitalization Rate: Respiratory	28.39	17.05	28.41	12.78

county nonattainment designation induced the state containing that county to submit a State Implementation Plan (SIP) detailing a strategy to achieve the NAAQS by a target date, and all counties not designated as nonattainment are considered in attainment and not required to produce a SIP. The initial target date for nonattainment areas to achieve the NAAQS was the end of 1994, but subsequent determination that some areas could not realistically attain the NAAQS in this time frame prompted the EPA to designate some areas as "serious" nonattainment and extend the target date to the end of 2001. For the purposes of our analysis, we consider the "initial" nonattainment designations to consist of any such designation that occurred between the years 1991-1995, as the distinguishing features between areas designated in 1991 and those in the next few years are thought to relate more to procedural issues and availability of data, not to air quality per se.

Potential Outcomes Approach and Causal Effects of Interest

The overall goal is to estimate causal effects of the initial nonattainment designations on Medicare outcomes in 2001. Importantly, the salient question is not "Did air pollution and health outcomes change during the time following the nonattainment designations?", but rather "Are air quality and health outcomes different after the nonattainment designations than they would have been if the designations had not occurred?".

More formally, we define the causal effect of interest as the comparison between two sets of potential outcomes: those that would occur if areas were designated as nonattainment for PM_{10} in 1991-1995 and those that would occur if the nonattainment designations had not occurred. Note that we forego the use of potential outcomes notation in the analysis of this case study (more formal use of potential outcomes notation appears in the Case Study 2: Accountability Assessment of Power Plant Emissions Controls Section). We consider the comparison of potential outcomes only among the locations that were actually designated as nonattainment, that is, the estimand of interest is what is known in the causal inference literature as the average treatment effect on the treated (ATT). Thus, we refine the causal question slightly to be "Among areas that were designated as nonattainment in 1991-1995, what were the causal effects of these designations?"

As a secondary objective, we also aim to characterize the existence of the anticipated causal pathway whereby the nonattainment designations impact health "through" reducing PM_{10} in 1999-2001. Because PM_{10} in this con-

text is an intermediate outcome that is expected to simultaneously be affected by the nonattainment designations and have bearing on the Medicare health outcomes, standard regression adjustment for ambient PM_{10} in 1999-2001 will not permit estimation of causal effects (Rosenbaum, 1984). Accordingly, we use principal stratification (Frangakis and Rubin, 2002) to quantify the extent to which causal effects of the designations health outcomes are 1) *associative* with causal effects of the designations on PM_{10} versus 2) *dissociative* with causal effects of the designations on PM_{10} (see the "Causal Pathways" Analyses: Causal Mediation Analysis and Principal Stratification Section). An associative effect that is large relative to the dissociative effect indicates that the causal effect of the designations on Medicare health outcomes is greater in areas where pollution was causally reduced than in areas where there is little or no effect of the designations on pollution. Dissociative effects that are similar in magnitude to associative effects indicate that the health impact of the intervention is similar regardless of whether the intervention decreased ambient pollution, which suggests the presence of other causal pathways through which the designations impact health without changing average ambient PM_{10} during 1999-2001. Recall that causal reductions in pollution are defined based on whether pollution was lower with the nonattainment designation than it would have been without the designation, not based on whether pollution decreased across time.

As outlined in the Statistical Perspectives for Causal Accountability Assessment Section, these types of causal questions can be framed as a hypothetical two-armed experiment with one "intervention" arm corresponding to the observed allocation of nonattainment designations and the other "control" arm corresponding to the hypothetical scenario with no nonattainment designations. Potential outcomes in nonattainment areas are observed under the intervention condition representing the actual nonattainment designations. Since potential outcomes are not observed under the hypothetical control condition without nonattainment designations, we use observed pollution and health outcomes in attainment areas that were not subject to SIP measures to characterize what would have happened in nonattainment areas had the nonattainment designations not occurred. Thus, the attainment areas can be construed as a "control group" for studying the effect of the nonattainment designations. The obvious threat to validity of estimating causal effects of the nonattainment designations by comparing outcomes with attainment areas is that the designations are decidedly not randomly assigned, meaning that attainment areas may share important differences with nonattainment areas, which is evident from Table 1. Using data on attainment areas to

learn about what would have happened in nonattainment areas requires careful confounding adjustment.

Estimation of Propensity Scores for Confounding Adjustment

The *propensity score* (Rosenbaum and Rubin, 1983; Stuart, 2010) is a nearly ubiquitous tool for adjusting for confounding to estimate causal effects with observational data that do not enjoy the benefits of randomization. The motivation for the propensity score in our setting is to construct groups of attainment and nonattainment locations that are comparable with respect to the covariates listed in Table 1 so that comparing outcomes in attainment and nonattainment areas does not suffer from confounding on the basis of these factors (see the Methods for Confounding Adjustment: Propensity Scores Section). The key assumption we adopt for confounding adjustment is that of *strong ignorability* that the covariates listed in Table 1 constitute (or are proxies for) all factors that could confound comparisons between attainment and nonattainment areas. This amounts to the familiar "no unmeasured confounding" assumption, and we revisit this assumption in the Conclusion and Discussion of Case Study 1 Section.

We use a logistic regression for the probability of a nonattainment designation to estimate the propensity score, with predicted probabilities from this model representing the estimated propensity scores. Covariates included in this propensity score model are those noted in Table 1. If the propensity score model is adequately specified and the factors in Table 1 comprise all of the relevant confounders (i.e., if there is no unmeasured confounding), then average outcomes in attainment areas represent what would have happened (on average) in nonattainment areas with similar values of the propensity score, just as would be the case in a randomized study.

Figure 5(a) depicts the distribution of estimated propensity scores for all 547 attainment and nonattainment locations. As expected, locations designated as nonattainment tend to have higher estimated propensity scores, but note that a wide range of estimated propensity scores are represented by both attainment and nonattainment areas. However, note in Figure 5(a) that locations with estimated propensity scores greater than 0.98 are exclusively nonattainment locations, with no attainment location having estimated propensity score in this range. This phenomenon, where areas of the estimated propensity score distribution only have representation from either the treated or the control group, is sometimes referred to as a lack of "overlap" between propensity score distributions

(Crump et al., 2009). Areas of the propensity score distribution that exhibit lack of overlap are not appropriate for making causal inferences, and observations lying in non-overlapping portions of the propensity score distribution should be removed or "pruned" from the analysis data set to prevent model-based extrapolation beyond the range of observed data (Ho et al., 2007; Crump et al., 2009).

In this instance, 49 nonattainment locations had estimated propensity scores greater than 0.98, that is, these observations did not "overlap" with estimated propensity scores from attainment areas. The implication of this lack of overlap is that each of these 49 nonattainment areas exhibits a constellation of the covariates in Table 1 that does not resemble that of *any* attainment area, leaving no observed information to learn about what would have happened in these areas had they not been designated. Put another way, these observations lack appropriate "control" observations. Analogously, 3 attainment areas had estimated propensity scores that did not overlap with those estimated in nonattainment areas. This lack of overlap is not surprising, as we would expect that, for example, there are areas of California's Central Valley having population demographics and pollution levels that do not resemble those of *any* other part of the country. For estimation of causal effects that do not rely on model-based extrapolation of the confounding adjustment, we discard the 52 observations without overlapping propensity scores to yield an pruned analysis sample that includes 495 monitoring locations, 219 of which lie in nonattainment areas (King and Zeng, 2006; Ho et al., 2007). This reduces the study population to 3,555,934 Medicare beneficiaries. See Figure 4(b) for the locations of the remaining monitoring locations. Appendix A.2 displays the locations of the discarded observations. Note that, strictly speaking, discarding such observations means that estimates from our analysis cannot technically be regarded as ATTs, as subsequent estimates only pertain to the subset of retained nonattainment lcoations.

After pruning the sample as described above, confounding adjustment is accomplished by grouping locations together based on the values of the estimated propensity score, as locations with similar values of the propensity score can be regarded as similar on the basis of all observed confounders. We classify the pruned analysis sample into 5 subgroups based on the quintiles of the estimated propensity score, each subgroup containing attainment and nonattainment locations that have similar values of the propensity score (i.e., are comparable with regard to the factors in Table 1). Table 2 lists the number of attainment and nonattainment areas in each propensity score

subgroup. Conditional on estimates of the propensity score, any models for pollution and health outcomes can be used to estimate causal effects in a manner that is much less susceptible to observed confounding (Ho et al., 2007).

Checking Covariate Balance

One essential benefit of propensity scores is the possibility to check the extent to which grouping observations based on estimated propensity scores "works" in the sense that it ensures the construction of groups of attainment and nonattainment locations that are in fact comparable on the bases of the factors in Table 1. If covariates are balanced between attainment and nonattainment areas within propensity score subclass, the potential for these covariates to confound the analysis of causal effects is greatly reduced. One common metric for checking covariate balance is the standardized difference between attainment and nonattainment observations (Stuart, 2010). This quantity can be calculated for each covariate as a way to summarize whether a covariate is in fact balanced between attainment and nonattainment areas, with values closer to zero indicating better average balance (and less susceptibility to confounding). We calculated standardized differences for each covariate before employing the propensity score, among all 547 monitoring locations in the original data set, and depict these values with the red line in Figure 6. These standardized differences in the unadjusted sample can be regarded as a measure of the potential for bias in causal effect estimates due to differences between attainment and nonattainment locations. We also calculate the standardized difference for each covariate within propensity score subclass as a measure of the covariate similarity among attainment and nonattainment locations with similar values of the propensity score. The blue line of Figure 6 plots, for each covariate, the average of this value across the 5 propensity score subclasses. The conclusion from Figure 6 is that, despite the stark differences between attainment and nonattainment locations in the entire sample (red line), the propensity score does an adequate job of balancing the covariates in Table 1 between attainment and nonattainment locations, within propensity score subclass (blue line).

Figure 5: Histograms of estimated propensity scores for attainment and nonattainment areas before and after pruning observations with non-overlapping propensity score estimates

(a) Full Monitor Set, 547 locations

(b) Pruned Monitor Set, 495 locations

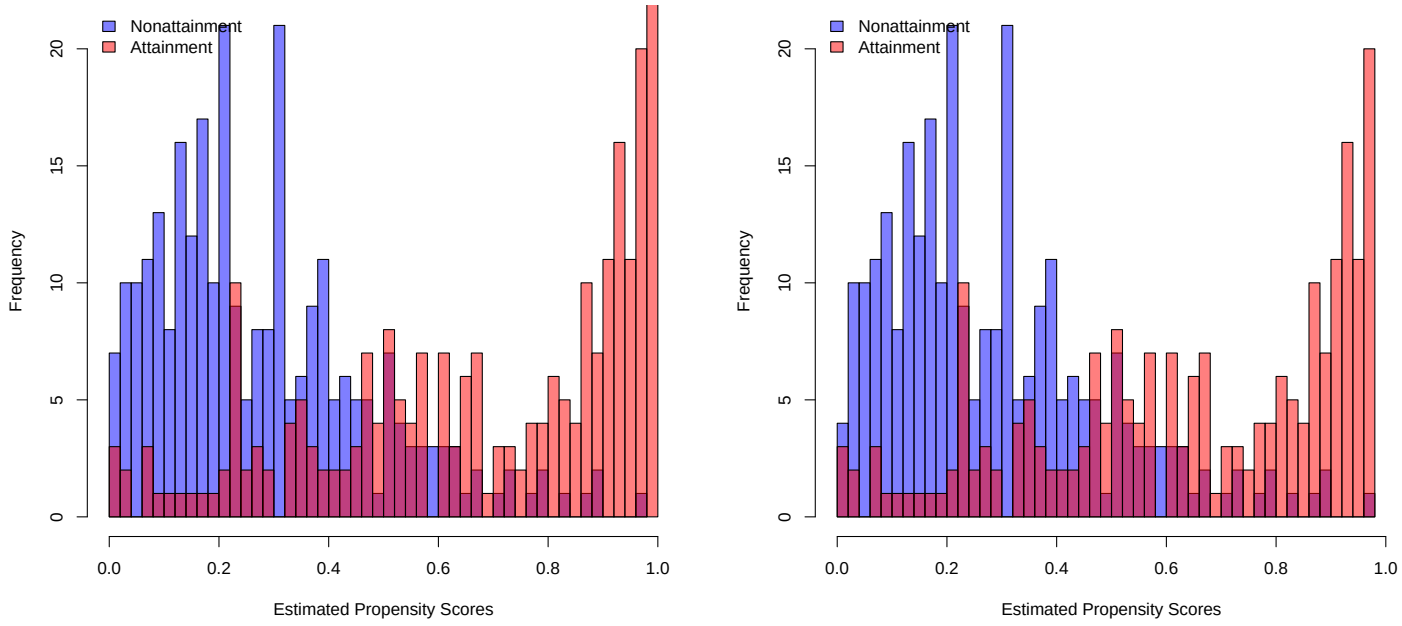
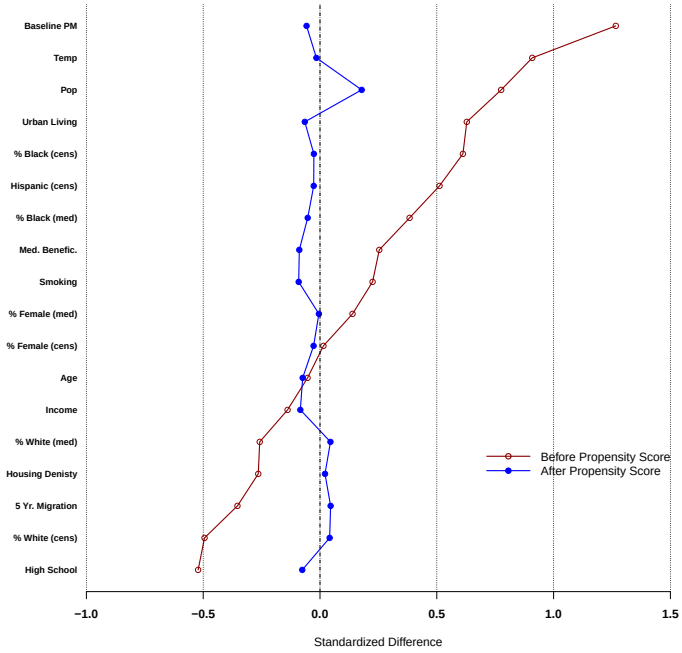


Table 2: Number of attainment and nonattainment areas in each of the five propensity score subclasses used for confounding adjustment

	Propensity Score Quintile				
	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}
Attainment	87	78	71	33	7
Nonattainment	12	21	28	66	92
Total	99	99	99	99	99

Figure 6: Description of covariate balance before (red line) and after (blue line) propensity score subclassifcation, as summarized by average standardized differences between nonattainment and attainment areas across each available covariate.



Models for Estimating Causal Effects

Note that none of the preceding discussion of propensity scores involves Medicare health outcomes, nor does it pertain to any particular statistical model for actually estimating causal effects. Rather, we have only formalized the causal effects of interest (ATTs), formulated the relevant "treatment" and "control" groups, and employed propensity scores to construct an analysis data set that will serve as the basis for estimating causal effects. Conditional on estimates of the propensity score, we used parametric models for potential outcomes under attainment and nonattainment designations to predict potential outcomes that are not observed, namely, the potential pollution and health outcomes that would have occurred in nonattainment areas had the designations never occurred. Insofar as these predictions can be regarded as an accurate reflection of what would have happened absent the designations, they can be used to estimate causal effects. Our analysis relies on two such models: 1) a spatial hierarchical regression model for (log-transformed) ambient PM_{10} concentrations during 1999-2001 and 2) log-linear Poisson regression models for each Medicare mortality and hospitalization outcomes. All regression models adjust for propensity score subclasses and also for individual variables from Table 1 to adjust for any residual confounding not accommodated by the propensity score model and to improve efficiency. See Zigler et al. (2012) and Appendix A for details of these model specifications, Markov chain Monte Carlo (MCMC) procedure used for estimation, and further technical detail.

Case Study 1 Results

Figure 7 depicts the annual average ambient PM_{10} for each of the 547 monitors for the years 1990 - 2001. As expected, ambient PM_{10} in the early 1990s tended to be higher in nonattainment areas, but both attainment and nonattainment areas had ambient PM_{10} below the NAAQS during this time frame. Also note that ambient average PM_{10} is decreasing similarly in both attainment and nonattainment areas.

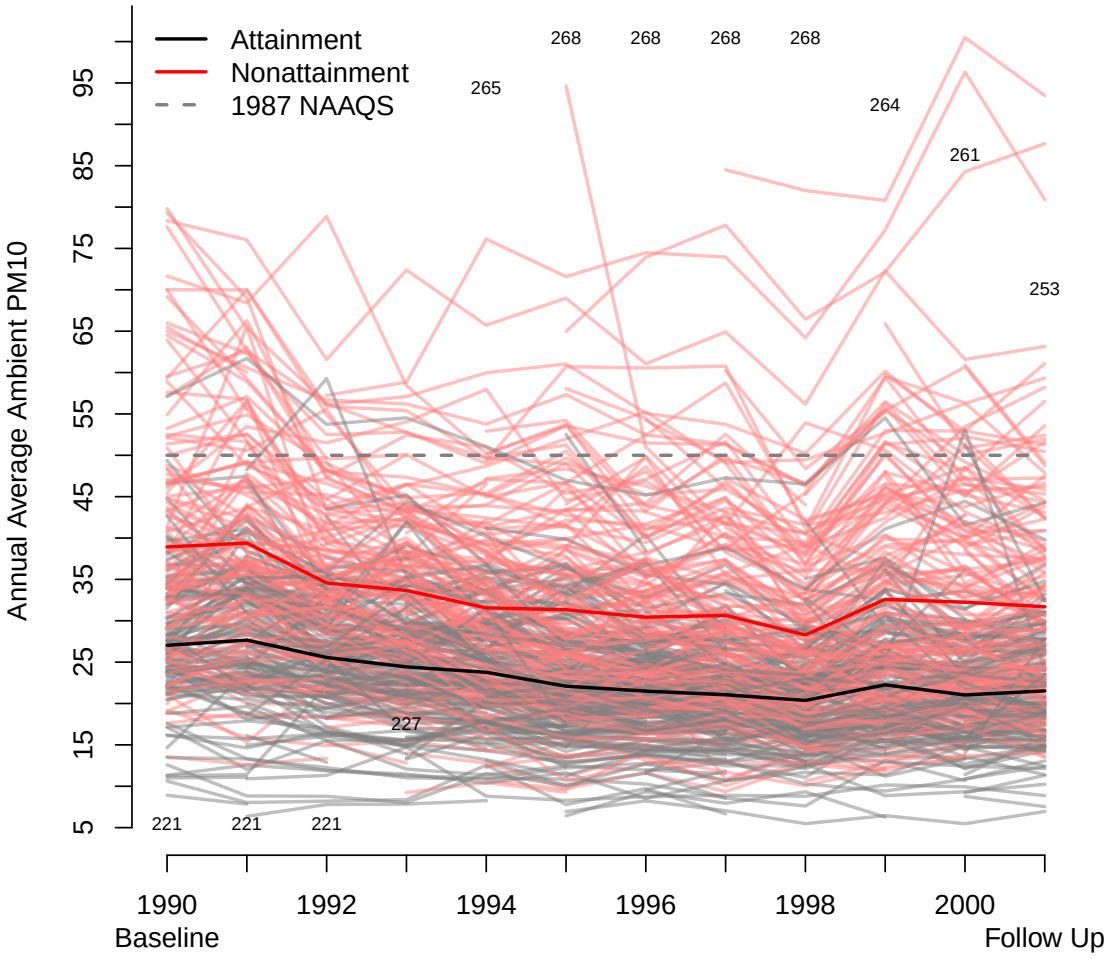
Unadjusted Comparisons

Using data on the entire sample of 547 monitoring locations indicates that, between the baseline and follow-up time periods, average ambient PM_{10} reduced by 8.8 $\mu g/m^3$ in nonattainment areas, from 40.4 in 1990 to 31.6 in 1999-2001. The analogous decrease in attainment areas was smaller: 5.4 $\mu g/m^3$, from 27.0 in 1990 to 21.6 in 1999-2001. The p-value comparing these changes from a two-sample t-test was $p < 0.001$. Among Medicare beneficiaries residing near one of the 547 monitors, the average rate of all-cause mortality (per 1000 person-years) in 2001 was similar in nonattainment and attainment areas (62.5 vs. 62.6, p-value from 2 sample t-test $= 0.952$). The average rate of CVD-related hospitalizations (per 1000 person-years) in 2001 was higher in nonattainment areas (92.1 vs. 83.7, p-value from 2 sample t-test $< .001$). Average rates of respiratory-related hospitalizations were similar in nonattainment areas and attainment areas (28.4 vs. 28.4, $p = 0.991$). Note that these unadjusted comparisons are likely confounded due to differences between attainment and nonattainment areas.

Average Causal Effects on Average Annual Ambient PM_{10} during 1999-2001

Using the propensity score approach outlined above and confining interest to the ATT among the 219 nonattainment areas in the pruned sample, we estimate the causal effect of the nonattainment designations on average ambient PM_{10} during 1999-2001 using the spatial hierarchical model outlined in Appendix A.3, adjusted for the propensity score subclass and the variables in Table 1. The estimated causal effect of the nonattainment designations on average ambient PM_{10} during 1999-2001 is -1.17 $\mu g/m^3$, with 95% posterior interval (-7.33, 4.00). This indicates that, among these 219 areas, average ambient PM_{10} during 1999-2001 was slightly lower than it would have been if the nonattainment designations had not occurred, that is, there is some evidence that the nonattainment designations had a causal effect on 3-year average ambient PM_{10} during 1999-2001. This highlights the likely possibility that decreases in this measure of PM_{10} during this time frame (evident from Figure 7) are likely due in part to factors affecting both attainment and nonattainment areas.

Figure 7: Trends in annual average ambient PM_{10} : 1990 - 2001. Thin lines represent individual monitoring locations, thick lines represent the average across all locations. The number listed for each year is the total number of nonattainment areas in that year.



Average Causal Effects on Medicare Health Outcomes

For the Medicare health outcomes, we used the propensity score approach outlined above and the models outlined in Zigler et al. (2012) and Appendix A.3 to estimate ATTs among the 219 nonattainment locations in the pruned sample. Models used for predicting potential outcomes adjust for propensity score subclass and the variables indicated in Table 1. Figure 8 summarizes posterior distributions of the average causal effects of the nonattainment designations on Medicare mortality, CVD-related hospitalization, and respiratory hospitalization rates (per 1000 beneficiaries). For all-cause mortality, the posterior mean ATT was -1.08 deaths per 1000 Medicare beneficiaries, with 95% posterior interval (-3.27, 0.99), suggesting that the nonattainment designations caused a decrease in mortality (i.e., that the average mortality rate in nonattainment areas was 1.08/1000 lower in 2001 than it would have been had these areas not been designated nonattainment). For CVD hospitalizations, the posterior mean ATT was 1.44 hospitalizations per 1000 person-years, with 95% posterior interval (-4.64, 7.16), which does not indicate any causal effect on CVD-related hospitalizations among Medicare beneficiaries relative to what would have occurred without the designations. For respiratory hospitalizations, the posterior mean (95% interval) average causal effect of the nonattainment designations was -1.47 (-3.86, 0.70) hospitalizations per 1000 person years, indicating that the nonattainment designations causally reduced respiratory-related hospitalizations, relative to what would have occurred without the designations. Note that all of the ATT estimates had 95% uncertainty intervals that include 0, and that the analysis of CVD-related hospitalizations exhibits the most uncertainty.

Associative and Dissociative Effects

To provide some insight into the existence of the anticipated causal pathway whereby the nonattainment designations decrease PM_{10} which, in turn, causes improvement in Medicare health outcomes, we use the same models described above (conditional on the propensity score and covariates in Table 1) to estimate associative and dissociative effects. Dissociative effects in this context are the causal effects of nonattainment designations on health outcomes among locations that are estimated to have experienced little or no causal effect on ambient PM_{10} during 1999-2001, i.e., where the estimated causal effect on this measure of PM_{10} is less than 5 $\mu g/m^3$. Associative

effects in this context are the causal effects of nonattainment designations on health outcomes among locations where ambient PM_{10} during 1999-2001 is estimated to have decreased substantially as a result of the designation, i.e., where the estimated causal effect on this measure of PM_{10} is a decrease of at least 5 $\mu g/m^3$.

For the mortality outcome, the posterior mean (95% interval) average dissociative effect is -1.90 (-5.52,1.87) deaths per 1000 beneficiaries, providing some evidence that mortality was reduced even in areas where PM_{10} during 1999-2001 was not causally affected. The posterior mean (95% interval) associative effect is -0.46 (-4.03, 2.64), indicating no evidence of a causal effect on mortality in locations where PM_{10} is estimated to have been causally decreased by more than 5 $\mu g/m^3$.

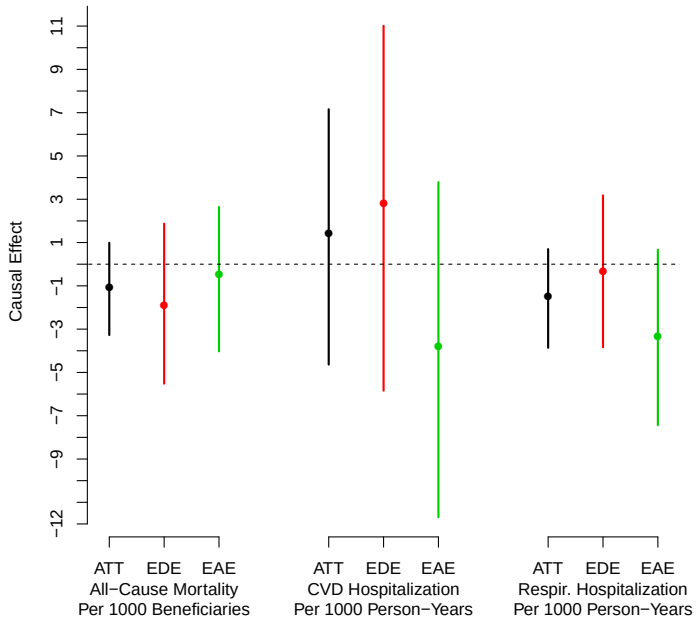
For the CVD-related hospitalization outcome, the posterior mean (95% interval) average dissociative effect is 2.83 (-5.84, 11.01) hospitalizations per 1000 person-years, providing no evidence that CVD hospitalizations were causally affected in areas where PM_{10} during 1999-2001 was not causally affected, with a point estimate suggestive of an increase in CVD hospitalizations in these areas. The posterior mean (95% interval) associative effect is -3.78 (-11.69, 3.79), which provides little evidence of a causal reduction in CVD hospitalization in locations where PM_{10} is estimated to have been causally decreased by more than 5 $\mu g/m^3$. Note the wide uncertainty intervals for both dissociative and associative effects for CVD-related hospitalizations.

For the respiratory-related hospitalization outcome, the posterior mean (95% interval) average dissociative effect is -0.31 (-3.84, 3.18), hospitalizations per 1000 person-years, providing no evidence that respiratory hospitalizations were causally affected in areas where PM_{10} during 1999-2001 was not causally affected. In contrast, the posterior mean (95% interval) associative effect is -3.34 (-7.43, 0.67), indicating a causal reduction in respiratory hospitalizations in locations where PM_{10} is estimated to have been causally decreased by more than 5 $\mu g/m^3$. Among the three outcomes, only the analysis of respiratory hospitalizations indicates an associative effect that is larger in magnitude than the dissociative effect, which is suggestive of the anticipated causal pathway; respiratory hospitalizations are not estimated to have been affected in areas where PM_{10} was not substantially affected, and these hospitalizations are estimated to have been causally reduced in areas where PM_{10} was causally reduced relative to what would have occurred without the nonattainment designations.

Table 3: Causal Effect Estimates for overall, associative, and dissociative effects in the analysis of PM_{10} nonattainment designations.

	Overal Average Causal Effect			Average Dissociative Effect			Average Associative Effect		
	Mean	2.5%	97.5%	Mean	2.5%	97.5%	Mean	2.5%	97.5%
Ambient PM10	-1.17	-7.33	4						
All-Cause Mortality	-1.08	-3.27	0.99	-1.9	-5.52	1.87	-0.46	-4.03	2.64
CVD Hospitalization	1.44	-4.64	7.16	2.83	-5.84	11.01	-3.78	-11.69	3.79
Respiratory Hospitalization	-1.47	-3.86	0.7	-0.31	-3.84	3.18	-3.34	-7.43	0.67

Figure 8: Posterior mean point estimates and 95% posterior probability intervals for overall, associative, and dissociative effects in the analysis of PM_{10} nonattainment designations.



Conclusion and Discussion of Case Study 1

We have employed the principles of the causal inference perspective described in the Statistical Perspectives for Causal Accountability Assessment Section to provide the first direct accountability assessment of one key feature of US air pollution regulatory policy: the initial PM_{10} nonattainment designations that followed from the 1990 CAA Amendments. Using a potential outcomes perspective, we explicitly define and estimate the causal effects of this specific set of regulatory decisions. Although ambient PM_{10} decreased in both attainment and nonattainment areas during the time frame of study, our results provide some evidence that 3-year average ambient PM_{10} during 1999-2001 among areas designated as nonattainment in 1991-1995 was lower due to the nonattainment designations than it would have been if the designations had never occurred. However, the small magnitude of this effect and associated uncertainty highlights the likely possibility that either the reductions in ambient PM_{10} observed between the early 1990s and early 2000s are attributable to other control measures that exist outside of the paradigm of nonattainment designations, or simply that the average ambient concentration during 1999-2001 is not an appropriate measure to capture the impact of the nonattainment designations. Furthermore, many areas likely started to take action to reduce air quality before 1990 in anticipation of the impending designations, which is not reflected in our analysis that relies on data dating back only to 1990. Despite the modest effect on average ambient PM_{10} during 1999-2001, our results provide evidence that the nonattainment designations causally reduced mortality and respiratory-related hospitalizations among Medicare beneficiaries residing near a monitor located in a nonattainment area, as compared to what would have occurred if the nonattainment designations had not taken place.

Results from our investigation of the presumed causal pathway whereby nonattainment designations improve health outcomes through reducing ambient PM_{10} differed depending on the outcome of interest. The principal stratification analysis of the respiratory-related hospitalization outcome indicated an associative effect that was much larger in magnitude than the dissociative effect. The estimated dissociative effect near zero indicates that the nonattainment designations did not cause reductions in hospitalizations among areas where PM_{10} was not substantially causally affected. The estimated associative effect that is different from zero indicates a causal

reduction in respiratory hospitalizations in areas where PM_{10} is estimated to have been causally reduced by more than 5 $\mu g/m^3$. Knowledge that respiratory hospitalizations were causally reduced only when PM_{10} was also causally reduced (large associative effect relative to dissociative effect) suggests the anticipated causal pathway, although this principal stratification analysis can not conclusively indicate the improvement in health outcomes is due to the causal reduction in PM_{10} , as would be the case in a formal mediation analysis (VanderWeele, 2009). In particular, our analysis cannot rule out the possibility that the correspondence between effects on health and effects on pollution is driven by a factor other than the nonattainment designations, nor can our analysis discern whether health effects are due to an alternative causal pathway present within the areas where ambient PM_{10} during 1999-2001 was causally reduced (see the Case Study 2: Accountability Assessment of Power Plant Emissions Controls Section for a formal causal mediation analysis). Nonetheless, we argue that knowledge that the causal effect of PM_{10} nonattainment designations on respiratory hospitalization outcomes is most pronounced in areas exhibiting causal reductions in ambient PM_{10} represents a result that is useful for informing future policies.

The principal stratification analysis of the CVD-related hospitalization outcome showed a similar pattern, with causal effects on this outcome most strongly pronounced in areas where the nonattainment designations decreased PM_{10} , however all estimates for this outcome were subject to substantial uncertainty. The principal stratification analysis of the mortality outcome provided no evidence that dissociative effects were different from associative effects, which suggests that any evident causal effects of nonattainment designations on mortality are likely due to causal pathways other than the impact on 3-year average ambient PM_{10} in 1999-2001. Examples of other important causal pathways include simultaneous or synergistic impacts on other pollutants, the initiation of economic consequences that affect health outcomes in the long term, or even other measures of PM_{10} , possibly in different time frame (HEI Accountability Working Group, 2003).

As with virtually all air pollution epidemiology, one significant challenge to causal inference is the prospect of confounding. In the accountability context considered here, confounders are factors that differ between attainment and nonattainment areas that also bear some relationship with pollution and/or health outcomes. Note that this is slightly different from studies of exposure-response relationships, where confounders are generally considered to be factors jointly associated with pollution and health outcomes. Observed confounding in our context is

particularly pronounced, as nonattainment areas were designated precisely because they exhibited (or contributed to) poor air quality, which is associated with a multitude of factors that differentiate attainment and nonattainment areas. Our propensity score strategy was able to group together attainment and nonattainment locations that were similar on the basis of baseline pollution levels, characteristics of the Medicare beneficiaries residing within 6 miles, and numerous features of the general nearby population, thus minimizing the chance of confounding with regard to these factors. Importantly, we discarded monitor locations in nonattainment areas that were not similar to *any* attainment area on the basis of observed confounders (and vice versa). Including these locations in our analysis would estimate causal effects that necessarily rely on model extrapolation beyond the information contained in the observed data, while removing them restricts our conclusions to only a subset of nonattainment areas. While our strategy is specifically designed to mitigate biased due to measured confounders, the prospect of unmeasured confounding remains a threat to the validity of our results. If there exist unmeasured factors related to pollution and/or health outcomes that, even after adjustment for all observed factors in Table 1, still differ between attainment and nonattainmnent areas, then our results are subject to unmeasured confounding.

One consequence of framing our approach in a formal potential-outcomes framework is the need to precisely define a specific intervention of interest, here, the nonattainment designations. While we argue that defining and estimating causal effects of the nonattainment designations is of considerable value, it should be noted that the designations resulted in a wide variety of specific actions to control air quality on state and local levels, and sometimes resulted in no action at all, for example, when regional control strategies were expected to reduce PM_{10} to achieve the NAAQS regardless of the nonattainment designations. This diversity of action (or inaction) is one likely explanation for the fact that annual ambient PM_{10} decreased similarly in both nonattainment and attainment areas. In this sense, estimating the causal effects of nonattainment designations is akin to an "intention to treat" analysis in clinical studies that consider causal effects of assignment to an intervention, as opposed to actual receipt of that intervention. Available data on specific actions taken on local scales (e.g., measures in a SIP) could facilitate a causal analysis of these actions, but the use for informing future policy could be limited because specific control actions are often highly specialized to local circumstances (e.g., spraying wind blown dust in Central California) and may not be replicable or relevant in other areas or at future time points. In focusing on the

causal effects of a set of federal-level regulatory decisions, we exchange some level of detail with regard to actual control measures for precision in definition of the intervention and evidence of the effectiveness of a regulatory process that can (and will) be replicated in the future.

A related limitation of our analysis is that it does not explicitly account for the regional nature of air quality control. Due to issues such as regional pollution transport, actions undertaken in SIPs in particular areas may have impacts that spread across to other areas. The prospect of actions in nonattainment areas having effects that spill over into nearby attainment areas would dilute the causal effects we wish to estimate; pollution and health outcomes in attainment areas are likely improved as well, and the present analysis does not account for this improvement. Thus, pollution and health outcomes under a setting where nonattainment designations had never occurred may have actually been worse than was indicated in the "control group" of our analysis that may have experienced benefits of the designations. This phenomenon is known in the causal inference literature as *interference*. Our work in Zigler et al. (2012) outlines an assumption about interference under which our approach would be robust to the effects of regional pollution transport. The assumption relies on the feature that nonattainment designations implicitly considered regional transport to some extent in that a particular area's designation could be based on contributions to air quality in other areas.

Our analysis makes use of a vast data resource that links together information on regulatory actions, ambient air quality, population characteristics, and health information on the entire US Medicare enrollment population. Accountability assessment of large-scale regulatory interventions relies on the large-scale availability of health data such as administrative hospital claims available on millions of individuals. Our analysis focused on the susceptible population of elderly in the US that lived near a PM_{10} monitor in the EPA monitoring network. Alternative pollution measurement techniques (e.g., satellite measurements or spatial extrapolation) could expand such an analysis to consider individuals not residing close to a monitor location, and other administrative data sources (e.g., emergency department or electronic health records) could be used to focus on other populations.

Nonattainment designations are one key mechanism for air quality control in the US, and represent a key step to the enforcement of the NAAQS. While our analysis entails important limitations, it provides important evidence of the effectiveness of this integral feature of air quality management in the US, and represents a distinct

perspective that should be interpreted in conjuncture with - not instead of - the large body of epidemiological research motivating the setting and enforcement of NAAQS.

Additional work in Progress

In addition to the case study presented above, we have engaged in a variety of related research endeavors designed to improve direct accountability assessment of the PM_{10} nonattainment designations. In Cefalu and Zigler (2015), we propose new methodology that generalizes the approach of omitting observations that lack propensity score "overlap" to allow for a stochastic filtering that ultimately weights causal estimates according to posterior evidence that each observation has a comparable observation in the opposite treatment group. We are working to deploy this methodology to analyze the causal effects of nonattainment designations over time. We are also working to corroborate the results of the analysis presented here with newly developed methods for Bayesian nonparametric principal stratification and causal mediation analysis that will be presented in the multipollutant context in the Case Study 2: Accountability Assessment of Power Plant Emissions Controls Section.

Case Study 2: Accountability Assessment of Power Plant Emissions Controls

Here we present an analysis of the second case study, which investigates the causal impacts (on multiple emissions and on ambient $PM_{2.5}$) of SO_2 scrubbers on coal-fired power plants. The primary focus of this case study is to illustrate the ideas outlined in the "Causal Pathways" Analyses: Causal Mediation Analysis and Principal Stratification Section and our newly-developed methods for principal stratification and causal mediation analysis in the multipollutant setting. Accordingly, the discussion in this section contains significantly more technical methodlogical detail than the previous sections.

Title IV of the Clean Air Act established the Acid Rain Program (ARP), which required major emissions reductions of SO_2 (and other emissions) from American power plants. The goal of this program was to reduce total SO_2 emissions by ten million tons relative to 1980 levels (29.5 million tons per year). This drop was to be achieved mostly through cutting emissions from electricity-generating units (EGUs), a process enacted in two

stages. Phase I (1995-1999) required 263 extremely polluting units to significantly reduce their emissions. Phase II, which began in 2000, placed a target SO_2 emissions cap at 8.95 billion tons per year on about 3,200 EGUs, which cut power-sector emissions nearly in half from 1980 levels.

Impacts of the ARP have been evaluated extensively, and the program is generally lauded as a success story due to marked national decreases in SO_2 and NO_x coming at relatively low cost. Despite a 25% increase in electricity production over the first 14 years of the program, SO_2 emissions fell by 36% (U.S. Environmental Protection Agency 2011b), and the program met its long-term goal of reducing EGU annual SO_2 emissions to 8.95 tons by 2007, with emissions declining further through at least 2010 (Schmalensee and Stavins, 2012). Estimates of the annualized human health benefits of the entire ARP range from $50 billion to $100 billion (Burtraw et al., 1998; Burtraw, 1999; Chestnut and Mills, 2005; Banzhaf et al., 2006; Schmalensee and Stavins, 2012). Recent analyses have used air quality model simulations to provide more targeted estimates of the health benefits attributable to emissions reductions from EGUs (Levy et al., 2007, 2009; Buonocore et al., 2014). Whether attempting to quantify the health impacts of the ARP as a whole or provide analyses of emissions reductions from specific EGUs, the existing evidence of the health benefits relies heavily on presumed relationships between power plant emissions, ambient $PM_{2.5}$, and human health.

While power plants under the ARP had latitude to elect a variety of strategies to reduce emissions such as changes in combustion technology or shifts in fuel composition, one key strategy is the installation of a flue-gas desulfurization technology (henceforth, a "scrubber") to reduce SO_2 emissions. The precise extent to which installation of a scrubber reduces ambient $PM_{2.5}$ through reducing SO_2 emissions remains unknown, and has never been estimated empirically amid the realities of actual regulatory implementation where pollution controls may impact a variety of factors that are also related to the formation of $PM_{2.5}$.

The goal of this case study is to examine the causal effects of installing a scrubber on a coal-fired power plant on the ambient concentration of ambient $PM_{2.5}$. In particular, we aim to quantify the contribution of the presumptive causal pathway that a scrubber reduces SO_2 emissions, which reduces ambient $PM_{2.5}$, relative to the importance of other causal pathways due to concurrent reductions (or co-benefits) in other emissions or other factors. Thus, the question will be formally framed as one of mediation analysis: To what extent is the causal

effect of a scrubber (the "treatment") on ambient $PM_{2.5}$ (the "outcome") mediated through reduced emissions of SO_2, NO_x and CO_2 (the "mediators")?

To answer these questions, we use the data sources described in the Publicly-Available Data and Reproducible Research for Accountability Assessment Section to provide a more refined direct accountability assessment of the extent to which a particular emissions-control action reduces emissions and causes improvements to ambient air quality. Specifically, we evaluate the extent to which an SO_2 scrubber on a coal-fired power plant 1) causally affects emissions of SO_2 , NO_x , and CO_2 , 2) causally affects ambient $PM_{2.5}$, and 3) causes effects on ambient $PM_{2.5}$ in a manner that is mediated through reducing SO_2 , NO_x , and/or CO_2 . We focus in particular on the question of mediated effects to provide rigorous evidence of the presumed relationships between actions, emissions, and ambient pollution that found a great deal of existing health benefits analyses. Towards this end, we deploy our newly-developed Bayesian nonparametric statistical methods that draw on two frameworks for estimating causal effects in the presence of intermediate mediating variables: (1) principal stratification (Frangakis and Rubin, 2002) and (2) direct and indirect effects, or so-called "causal mediation analysis" (Robins and Greenland, 1992; Pearl, 2001; VanderWeele and Vansteelandt, 2014). Both frameworks require development of new statistical methods to accommodate the multipollutant nature of the problem.

Linked Data Sources

Using the tools described in the Publicly-Available Data and Reproducible Research for Accountability Assessment Section, we assembled a national database to conduct the investigation. We obtain annual emissions data on 269 coal-fired power plants during the year 2005. We also obtain various characteristics of the plants including NO_x emissions controls, annual average heat input in 2004, participating phase of the ARP (I or II), operating time in 2004, and coal sulfur content in 2004 (see Table 4).

We link each power plant to all ambient $PM_{2.5}$ monitors located within a 150km radius of the plant. Monitors located within 150km of more than one plant are linked to the closest plant. For each plant, we calculate the average annual ambient $PM_{2.5}$ concentration in 2005 among all $PM_{2.5}$ monitors linked to that plant, as well as average temperature and barometric pressure during 2004. Constructing the data set in this way entails many

important limitations, mostly due to the realities of regional pollution transport. We clarify the causal quantities being estimated in the Defining the Intervention: SO_2 Scrubber Installation Section, and revisit this limitation in the Conclusion and Discussion of Case Study 2 Section.

Table 4: Summary statistics for covariates and outcomes available for the analysis of SO_2 scrubbers.

	Have scrubbers (n=70)		Have no scubber (n=199)	
	Mean	SD	Mean	SD
Monitor Data				
Average Ambient $PM_{2.5}$ 2005	11.59	3.88	13.14	2.68
Average Temperature 2004	13.05	4.73	13.74	3.57
Average Barometric Pressure 2004	718.45	52.71	741.12	24.20
Power Plant Level Data				
Total SO_2 Emission 2005	1320.03	1823.4	2222.26	2540.64
Total NO_x Emission 2005	917.75	774.03	605.38	557.11
Total CO_2 Emission 2005	565349	475743	391091.9	375841.8
Unit Level Data				
NO_x Scrubber Installation, Jan. 2005	0.89	0.3	0.79	0.36
Average Heat Input 2004	5501996	4623661	3845140	3683800
Phase II Indicator 2004	0.83	0.36	0.76	0.41
Total Operating Time 2004	7697.35	650.73	7371.97	1007.99
Sulfur Content in Coal 2004	1.13	0.89	0.65	0.43

Defining the Intervention: SO_2 Scrubber Installation

A power plant (or power generating facility) can consist of multiple EGUs, each of which may or many not be equipped with a scrubber. Thus, to define the intervention at the level of the facility, we regard any facility to be "treated" with a scrubber if at least 10% of the total heat input for that facility can be attributed to EGUs within that facility that have scrubbers installed as of January 2005. Appendix B.1 shows the distribution of % heat input from an EGU with a scrubber across all 269 power plants, which illustrates that the vast majority of facilities had nearly all or nearly none of their heat input attributed to EGUs with scrubbers. This suggests the relative unimportance of exactly how to define facilities as having been "treated" with a scrubber.

The "control" condition used for comparison is the setting where no scrubbers were installed. Thus, causal

effects in this case study relate to comparisons between emissions and ambient $PM_{2.5}$ that would be potentially observed if a particular facility did or did not adopt scrubbers to control SO_2 emissions. The "intervention group" of the study consists of the 70 power plants that had scrubbers in January 2005, and the "control group" consists of the 199 plants that did not. Note that three power plants were excluded from this analysis because they installed scrubbers during 2005.

Defining Potential Outcomes for Principal Stratification and Causal Mediation Analysis

The primary objective of this case study is to characterize the extent to which installing a scrubber impacts ambient $PM_{2.5}$, "through" reducing emissions of SO_2 , NO_x , and CO_2 . The "Causal Pathways" Analyses: Causal Mediation Analysis and Principal Stratification Section outlines the rationale behind two analytic perspectives on such questions. The Case Study 1: Accountability Assessment of PM_{10} Nonattainment Designations Section illustrates one application of principal stratification for accountability assessment in the context of a single pollutant. For the current case study, we refine the previous descriptions of potential outcomes, principal stratification, direct, and indirect effects to illustrate our statistical methods development to accommodate multipollutant accountability settings.

We formulate the approach with explicit potential-outcomes notation (Rubin, 1978). Consider a single power plant and let $Z \in \{0,1\}$ denote whether the power plant has scrubber(s) installed in January 2005, with $Z = 1$ denoting the presence of a scrubber. Let $\{M_k(z); k = 1,\ldots,K\}$ denote the potential emissions of K pollutants that would occur if the power plant were to have scrubber status $Z = z$, for $z = 0,1$. Henceforth, we fix $K = 3$ so that $M_k(z), k = 1,2,3$ denotes the potential emissions of SO_2 , NO_x , and CO_2 , respectively. The causal effect of the scrubber on emission k can then be defined as a comparison between $M_k(1)$ and $M_k(0)$ comparing emissions that would be observed under the "treatment" and "control" condition. Let $\mathbf{M}(z_1,z_2,z_3) \equiv \{M_1(z_1),M_2(z_2),M_3(z_3)\}$ denote potential emissions under a set of three scrubber statuses $\{z_1,z_2,z_3\}$. For example, $\mathbf{M}(1,0,0)$ would represent the potential SO_2 concentration under installation of a scrubber and the potential NO_x and CO_2 concentrations that would be observed if the scrubber had not been installed.

We similarly define potential $PM_{2.5}$ outcomes, but extend the notation to define potential concentrations under

different potential values of scrubber status, Z, and different possible values of emissions, $\mathbf{M}(z_1, z_2, z_3)$. Thus, in full generality, each power plant has a set of $2^4 = 16$ potential outcomes for $PM_{2.5}$, $Y(z; \mathbf{M}(z_1, z_2, z_3))$, which denote potential values of $PM_{2.5}$ that would be observed under intervention $Z = z$ with pollutant emissions set at values under interventions z_1, z_2, z_3. Definition of all 16 potential $PM_{2.5}$ concentrations implies that each emission could, at least in theory, be intervened upon independently $PM_{2.5}$ and the other emissions. Thus, it is worth noting that values of $Y(z; \mathbf{M}(z_1, z_2, z_3))$, can be categorized into two groups. For $z = z_1 = z_2 = z_3$, potential outcomes are observable from the data, that is, any power plant with a scrubber will have $Y(1; \mathbf{M}(1,1,1))$ observed, and any power plant without a scrubber will have $Y(0; \mathbf{M}(0,0,0))$ observed. Were refer to these as *observable* potential outcomes. In contrast, potential outcomes defined under any other values of the vector (z, z_1, z_2, z_3) represent potential outcomes where a power plant is simultaneously subject to different interventions, and can never be observed in practice. For example, $Y(1; \mathbf{M}(0,0,1))$ would represent the potential ambient $PM_{2.5}$ concentration near a plant under the hypothetical scenario where that plant installs a scrubber ($z = 1$), but where emissions of SO_2 and NO_x are set to what they would be without the scrubber ($z_1 = z_2 = 0$) and emissions of CO_2 are set to what they would be with the scrubber ($z_3 = 1$). We refer to these potential outcomes as *unobservable* (or *a priori* counterfactual), as they are never observed for any power plant. Estimating causal effects relying on such unobservable potential outcomes will rely on unverifiable assumptions relating each of these unobservable quantities to observed relationships in the data.

Note that the total effect (TE) of scrubber installation on ambient $PM_{2.5}$ can be defined as the comparison between the observable potential outcomes $Y(1; \mathbf{M}(1,1,1))$ and $Y(0; \mathbf{M}(0,0,0))$. Various other causal effects related to "causal pathways" will be defined based on comparisons between different combinations of the above potential outcomes.

Principal stratification defines causal effects based only on the observable potential outcomes $Y(1; \mathbf{M}(1,1,1))$ and $Y(0; \mathbf{M}(0,0,0))$. Associative effects represent causal effects of a scrubber on ambient $PM_{2.5}$ among power plants where emissions are causally affected by the scrubber. Dissociative effects represent causal effects of a scrubber on ambient $PM_{2.5}$ among power plants where emissions are not meaningfully affected by the scrubber. In the presence of multiple pollutants, associative and dissociative effects can be defined as functions of changes

in each of the $K = 3$ emissions. Following development in Zigler et al. (2012), we focus discussion on average (or expected) associative and dissociative effects defined as:

$$\begin{aligned}
\text{EDE}_{\mathcal{K}} &= E[Y(1;\mathbf{M}(1,1,1)) - Y(0;\mathbf{M}(0,0,0)) \,\big|\, |(\mathbf{M}(0,0,0) - \mathbf{M}(1,1,1))|_{\mathcal{K}} < C^D_{\mathcal{K}}], \\
\text{EAE}_{\mathcal{K}} &= E[Y(1;\mathbf{M}(1,1,1)) - Y(0;\mathbf{M}(0,0,0)) \,\big|\, |(\mathbf{M}(0,0,0) - \mathbf{M}(1,1,1))|_{\mathcal{K}} > C^A_{\mathcal{K}}],
\end{aligned}$$

where $|(\mathbf{M}(1,1,1) - \mathbf{M}(0,0,0))|_{\mathcal{K}}$ denotes a vector of absolute differences between potential emissions of the subset of pollutants in the set $\mathcal{K}$, with $>$ and $<$ representing element-wise comparisons between vectors of mediators. For example, $\mathcal{K} = \{1,2\}$ would be used to define associative and dissociative effects based only on causal effects on emissions of SO_2 and NO_x , without regard to the effect on CO_2 . Here, $C^A_{\mathcal{K}}$ denotes a vector of thresholds beyond which a change in each pollutant emission in $\mathcal{K}$ is considered meaningful, while $C^D_{\mathcal{K}}$ is a vector of thresholds below which changes in these pollutant emissions are considered not meaningful. For example, with $\mathcal{K} = \{1,2,3\}$ and $C^A_{\mathcal{K}=\{1,2,3\}} \equiv \{C^A_1, C^A_2, C^A_3\}$, EAE could estimate the average causal effect on $PM_{2.5}$ among power plants for which scrubber installation causally affected emissions of SO_2 , NO_x , and CO_2 in excess of C^A_1, C^A_2, C^A_3, respectively. Estimates of EAE and EDE are useful summary measures of causal effects on $PM_{2.5}$, on average, when emissions change or do not change, but the relationship between causal effects on $PM_{2.5}$ and causal effects on emissions can vary as an entire surface describing effects on $PM_{2.5}$ for any particular values of $\mathbf{M}(0)$ and $\mathbf{M}(1)$. In addition to estimating EDE and EAE for different $\mathcal{K}$ as defined above, we also estimate entire surfaces of, for example, how the causal effect on $PM_{2.5}$ varies as a function of the causal effect on each emission.

Causal mediation analysis relies on the definition of *natural direct* and *natural indirect effects* (Robins and Greenland, 1992; Pearl, 2001), which are defined based on potential outcomes that are described above as unobservable. Natural direct effects in this context are defined as causal effects of scrubber installation on $PM_{2.5}$ when emissions are set to their "natural" value that would be observed without a scrubber, thus representing the causal effect of scrubber installation on $PM_{2.5}$ that is "direct" in the sense that it is not attributable to changes in emissions. Formally, we define the natural direct effect (NDE) in the multipollutant setting as NDE =

$E[Y(1;\mathbf{M}(0,0,0)) - Y(0;\mathbf{M}(0,0,0))]$.

Natural indirect effects in this context are defined to represent causal effects of scrubber installation on $PM_{2.5}$ that are attributable only to emissions changes. In the multipollutant setting, different natural indirect effects are defined based on different multipollutant emissions. The joint natural indirect effect of all 3 mediators, $JNIE_{123}$ (i.e., the indirect effect attributable to changes in all three emissions), is derived by subtracting the natural direct effect from the total effect, $JNIE_{123} = TE - NDE = E[Y(1;\mathbf{M}(1,1,1)) - Y(1;\mathbf{M}(0,0,0))]$, where TE is as defined in the Defining Potential Outcomes for Principal Stratification and Causal Mediation Analysis Section.

In addition to $JNIE_{123}$, which is of interest, we introduce a decomposition of this joint effect into the natural indirect effect attributable to changes in different combinations of the $K = 3$ emissions. The $JNIE_{123}$ can be decomposed into emission-specific indirect effects and the joint indirect effects of all possible pairs of emissions. See Figure 10 for a graphical representation of the different natural indirect effects.

The *mediator-specific* NIE for the k^{th} emission is defined as a comparison between potential $PM_{2.5}$ outcomes where the k^{th} emission varies between what it would be with and without a scrubber, but all other emissions are fixed to the potential value that would be observed with the scrubber. The mediator-specific NIEs for emissions of SO_2 , NO_x , and CO_2 are define as:

$$NIE_1 = E[Y(1;M_1(1),M_2(1),M_3(1)) - Y(1;M_1(0),M_2(1),M_3(1))],$$

$$NIE_2 = E[Y(1;M_1(1),M_2(1),M_3(1)) - Y(1;M_1(1),M_2(0),M_3(1))],$$

and

$$NIE_3 = E[Y(1;M_1(1),M_2(1),M_3(1)) - Y(1;M_1(1),M_2(1),M_3(0))],$$

respectively.

In a similar fashion we can define the joint natural indirect effect attributable to changes in pairs of mediators j and k,$JNIE_{jk}$ (the second row in Figure 10). The joint natural indirect effect of mediators j and k are defined as differences between the potential $PM_{2.5}$ that would be observed with a scrubber and the analogous poten-

tial outcome but with pollutants j and k set to their values that would have been emitted without the scrubber. Specifically, the joint natural indirect effects for changes in SO_2 and NO_x is defined as:

$$\text{JNIE}_{12} = E[Y(1;M_1(1),M_2(1),M_3(1)) - Y(1;M_1(0),M_2(0),M_3(1))],$$

with the analogously-defined effects for (NO_x and CO_2) and (SO_2 and CO_2) defined as

$$\text{JNIE}_{23} = E[Y(1;M_1(1),M_2(1),M_3(1)) - Y(1;M_1(1),M_2(0),M_3(0))],$$

and

$$\text{JNIE}_{13} = E[Y(1;M_1(1),M_2(1),M_3(1)) - Y(1;M_1(0),M_2(1),M_3(0))],$$

respectively.

As can be seen in Figure 10, the joint natural indirect effect of each pair of mediators is not equal to the sum of corresponding mediator-specific NIEs. That is, our definitions of indirect effects do not assume additivity of effects, nor do they assume that the indirect effects are non-overlapping. This is essential in the multipollutant setting where multiple pollution emissions are measured contemporaneously and are not generally independent of each other. Thus, our methods development is in important contrast to the budding literature on mediation analysis with multiple mediators that tends to assume non-overlapping, independent, sequential, or additive effects (MacKinnon, 2008; Imai and Yamamoto, 2013; Daniel et al., 2014; VanderWeele and Vansteelandt, 2014).

Estimation: New Methods for Bayesian Nonparametric Mediation Analysis

Estimation and inference for the causal effects defined above are based on models for the joint distribution of outcomes and mediators,

$$[Y(0;\mathbf{M}(0,0,0)),Y(1;\mathbf{M}(1,1,1)),\mathbf{M}(0,0,0),\mathbf{M}(1,1,1)|\boldsymbol{X}],$$

where $\boldsymbol{X}$ denotes a vector of baseline covariates used to adjust for confounding (see Table 4). This joint distribution is not identified based on the observed data, as potential outcomes are never jointly observed in both the presence and absence of a scrubber. Thus, our estimation strategy can be characterized as consisting of three steps. First, we specify flexible Bayesian nonparametric models for the marginal distributions of observed data , which consist of values of $(Y(0;\mathbf{M}(0,0,0)),\mathbf{M}(0,0,0))$ observed for power plants that did not install scrubbers, and $(Y(1;\mathbf{M}(1,1,1)),\mathbf{M}(1,1,1))$ observed for power plants that did install scrubbers (Escobar and West, 1995; Mller et al., 1996; Jara et al., 2011). Second, we link all of these flexibly-modeled marginal distributions into a coherent joint distribution through the use of a Gaussian copula model (Nelsen, 1999). Unobserved (but observable) potential outcomes are then simulated from their posterior-predictive distributions to estimate the TE and the associative and dissociative effects. Finally, a series of assumptions are used to relate unobservable potential outcomes to observed relationships in the data to provide estimates of the natural direct and indirect effects. Details of the statistical models, the assumptions for identification, and the MCMC computational algorithm appear in Appendix B.

Case Study 2 Results

Table 4 indicates that the 150km area around power plants with SO_2 scrubbers installed in January 2005 had average ambient $PM_{2.5}$ that was lower, on average, than the areas surrounding power plants without scrubbers (11.6 vs. 13.1 $\mu g/m^3$). The power plants with SO_2 scrubbers also emitted less SO_2 , more NO_x , and more CO_2 than the plants without scrubbers, and they also tended to burn coal with a higher sulfur content.

Using the approach outlined above, having an SO_2 scrubber installed is estimated to cause SO_2 emissions to be -1.00 log(tons) lower, on average, than they would be without the scrubber (95 % posterior interval (-1.41, -0.64)). The analogous causal effects for NO_x and CO_2 emissions were 0.15 (-0.12, 0.41) and 0.06 (-0.11, 0.24), respectively, indicating that SO_2 scrubbers did not causally affect these emissions, on average. The total effect (TE) of having an SO_2 scrubber installed on ambient $PM_{2.5}$ within a 150km radius of a power plant is estimated to be -1.18 $\mu g/m^3$ (95% posterior interval -2.09, -0.29), indicating that having a scrubber installed causally reduces

ambient $PM_{2.5}$ relative to what would occur in the absence of a scrubber.

The principal stratification analysis can provide important evidence of the extent to which causal reductions in $PM_{2.5}$ are associative or dissociative with changes in emissions of SO_2 , NO_x , and CO_2 . We define average associative and dissociative effects for 1) the triplet of SO_2 , NO_x , and CO_2 emissions; 2) each pair of emissions, and 3) each emission individually. For the k^{th} emission, let σ_k denote the standard deviation of the estimated individual-level causal effect of a scrubber on the k^{th} emission, with $\sigma_1 = 1.526$, $\sigma_2 = 1.084$, $\sigma_1 = 0.945$. We consider changes within $0.5\sigma_k$ of 0 to represent little or no causal effect on emissions, and changes in excess of $0.5\sigma_k$ to represent meaningful changes in emissions. Thus, for the $EDE_{\mathcal{K}}$, we set $C^D_{\mathcal{K}} = 0.5\sigma_{\mathcal{K}}$, to define the average effect of a scrubber on ambient $PM_{2.5}$ among power plants where there is little or no estimated causal effect on the emissions in $\mathcal{K}$. For each $\mathcal{K}$, we define two types of EAE: EAE_1 representing the causal effect of a scrubber on $PM_{2.5}$ among power plants where emissions were *reduced* by at least $0.5\sigma_k$ relative to what they would have been without the scrubber, and EAE_2 representing the analogous effect among power plants where emissions were *increased* by at least $0.5\sigma_k$. These correspond to $C^A_{\mathcal{K}} = 0.5\sigma_{\mathcal{K}}$

Table 5: Posterior mean and 95% probability intervals for expected associative and dissociative effects of SO_2 scrubbers.

		SO_2	NO_x	CO_2	SO_2 & NO_x	SO_2 & CO_2	NO_x & CO_2	SO_2 & NO_x & CO_2
EAE_1	Mean	-1.402	-1.251	-1.469	-1.385	-1.605	-1.452	-1.549
	95% P.I.	(-2.323, -0.565)	(-2.197, -0.375)	(-2.458, -0.554)	(-2.348, -0.522)	(-2.606, -0.681)	(-2.418, -0.614)	(-2.531,-0.692)
EDE	Mean	-1.063	-1.253	-1.221	-1.047	-1.082	-1.245	-1.062
	95% P.I.	(-2.096, -0.062)	(-2.114, -0.358)	(-2.142, -0.349)	(-2.119, -0.030)	(-2.164, -0.065)	(-2.124, -0.360)	(-2.115,-0.067)
EAE_2	Mean	-0.669	-1.143	-0.985	-0.696	-0.643	-0.987	-0.675
	95% P.I.	(-2.066, 0.963)	(-2.210, -0.069)	(-1.988, -0.026)	(-2.192, 1.079)	(-2.197, 1.004)	(-2.057, 0.081)	(-2.250,1.012)

For each set of emissions, Figure 11 depicts the posterior mean estimate of EDE (blue circle), EAE_1 (red circle), and EAE_2 (grey circle). The size of the circle is proportional to the estimated proportion of power plants that contribute to the estimate, that is, the estimated proportion of plants that have changes in emissions in accordance with the values of $C^D_{\mathcal{K}}$ and $C^A_{\mathcal{K}}$. For example, the red circle in the first column of Figure 11 indicates that a scrubber is estimated to reduce SO_2 emissions by at least $C^A_{\mathcal{K}}$ for 58% of power plants, and the average causal effect on $PM_{2.5}$ among these power plants is a reduction of 1.40 $\mu g/m^3$ relative to what would have occurred if a scrubber had not been installed. The blue circle indicates that a scrubber is estimated to have little or no causal

effect on SO_2 emissions among 31% of power plants, and the average causal effect of the scrubber on ambient $PM_{2.5}$ in these plants is a reduction of 1.06 $\mu g/m^3$. The grey circle indicates that 11% of plants are estimated to cause increased emissions of SO_2 of at least $C^A_{\mathcal{K}}$, and the average causal effect of the scrubber on ambient $PM_{2.5}$ around these plants is a reduction of 0.67 $\mu g/m^3$. The second two columns of Figure 11 can be interpreted analogously, but for NO_x and CO_2 , respectively. Columns 4-6 of Figure 11 depict EDE, EAE_1, and EAE_2 defined based on changes in the corresponding pairs of emissions. For example, the red dot in column 5 of Figure 11 indicates that the average causal effect on $PM_{2.5}$ among power plants where both SO_2 and CO_2 were reduced was a reduction of 1.60 $\mu g/m^3$, and 20% of plants were estimated to experience such emissions reductions. Table 5 lists posterior mean and 95% posterior intervals of EDE, EAE_1, and EAE_2 for all possible $\mathcal{K}$. Overall, we see that all estimates of EDE, EAE_1 and EAE_2 are negative, implying that scrubbers causally reduce $PM_{2.5}$ regardless of the precise manner of emissions reductions of SO_2 , NO_x , and CO_2 . This indicates the existence of some other causal pathway whereby scrubbers reduce $PM_{2.5}$ in a manner that is not captured solely by the emissions reductions of these pollutants. Note that EAE_1 tends to be larger in magnitude than the EDE whenever $\mathcal{K}$ includes SO_2 emissions, implying that the scrubbers reduce $PM_{2.5}$ more when there is a larger reduction in SO_2 emissions. This is suggestive of the anticipated causal pathway whereby the scrubber reduces SO_2 which, in turn, reduces $PM_{2.5}$. A similar pattern is also slightly evident for reductions in CO_2 .

While estimates of EDE, EAE_1, and EAE_2 provide useful summary quantities of how the causal effect of a scrubber on $PM_{2.5}$ varies with the causal effect of the scrubber on emissions, we can also examine entire surfaces of how the scrubber effect on $PM_{2.5}$ varies with the causal effect on emissions. Figure 12 depicts 3-D surface plots for each emission, where the surface summarizes the estimated causal effect on $PM_{2.5}$ for any estimated causal effect of the scrubber on emissions. Figure 12(a) depicts the estimated effect on $PM_{2.5}$, $Y(1,\mathbf{M}(1,1,1)) - Y(0,\mathbf{M}(0,0,0))$, for any combination of $(M(0),M(1))$ across the range observed in the data. Note that nearly the entire surface lies below 0, indicating that the scrubber is estimated to reduce $PM_{2.5}$ regardless of the effect on SO_2 emissions. The surface is lowest in the region where $M(1) < M(0)$, and the steepest portion of the surface is in regions where $M(0)$ is high, that is, the causal effect on $PM_{2.5}$ is estimated to be most pronounced in power plants that would have the highest emissions absent a scrubber and exhibit a causal reduction in emissions due

to the scrubber. The analogous surfaces for NO_x and CO_2 in Figures 12(b) and 12(c) are much flatter, which suggests that any causal effect of a scrubber on $PM_{2.5}$ varies only slightly with varying causal effects on these emissions. Note that In Figure 12(a), the blue dots representing the estimated causal effect of the scrubber on SO_2 emissions for 1 MCMC iteration lie almost entirely in the region where $M(1) < M(0)$ reflecting that the scrubber is estimated to reduce SO_2 emissions for nearly all power plants. In Figures 12(b) and 12(c), the analogous blue dots follow more closely and symmetrically around the line $M(1) = M(0)$, reflecting that SO_2 scrubbers do not have strong effects on these emissions on average.

The overall conclusion of the principal stratification analysis is that 1) causal effects of scrubber on ambient $PM_{2.5}$ are evident regardless of the causal impacts on emissions of SO_2 , NO_x , and CO_2 ; 2) larger causal reductions in SO_2 are associated with larger causal reductions in $PM_{2.5}$, regardless of changes in other emissions, and 2) there is some evidence that scrubber effects on ambient $PM_{2.5}$ are also associated with reductions in CO_2 emissions relative to what they would have been absent the scrubber. While not conclusive about the mediated indirect effects whereby the scrubber reduces $PM_{2.5}$ through reducing SO_2 emissions, this analysis is highly consistent with the presence of such a causal pathway.

Augmenting the principal stratification analysis with assumptions about unobservable potential outcomes that conceive of independent manipulations of scrubbers and each emission individually permits estimation of natural direct and indirect effects. These effects speak more directly to the extent to which the effect of an SO_2 scrubber on ambient $PM_{2.5}$ within 150km is mediated through various emissions. Recall that the mediator-specific natural indirect effect of the k^{th} emission and the joint natural indirect effect of the j^{th} and k^{th} emissions are denoted as NIE_k and $JNIE_{jk}$, respectively, where $k = 1$ indicates SO_2, $k = 2$ indicates NO_x and $k = 3$ indicates CO_2.

Table 6 summarizes point estimates and 95% posterior intervals of the TE, NDE, JNE_{123}, $JNIE_{12}$, $JNIE_{23}$, $JNIE_{13}$, and the individual NIEs. Figure 13 depicts boxplots of the entire posterior distributions of these quantities. The posterior mean (95% posterior interval) NDE representing the direct effect of a scrubber on ambient $PM_{2.5}$ that is not mediated through any emissions changes is a reduction in ambient $PM_{2.5}$ of 0.74 (-1.67, 0.18) $\mu g/m^3$. This represents the amount that $PM_{2.5}$ would decrease if a scrubber were installed but emissions of SO_2 , NO_x , and CO_2 were somehow fixed to remain constant at what the would have been without the scrubber. The indirect

effect via all three emissions ($JNIE_{123}$) is estimated with posterior mean (95% posterior interval) -0.44 (-0.70, -0.23), which represents the reduction in $PM_{2.5}$ that would occur around a plant with a scrubber relative to what would happen if emissions of SO_2 , NO_x , and CO_2 were somehow changed to what they would have been absent the scrubber. The relative magnitudes of the TE and $JNIE_{123}$ indicate that over one third of the total effect of a scrubber on ambient $PM_{2.5}$ is jointly mediated through changes in SO_2 , NO_x , and CO_2 . The NIE for SO_2 (NIE_1) is estimated with posterior mean (95% posterior interval) -0.36 (-0.53, -0.22), indicating that most of the joint indirect effect is due to reductions in SO_2 . The NIE for NO_x (NIE_2) is also slightly negative (posterior mean (95% posterior interval) -0.10 (-0.26, 0.06)), and the NIE for CO_2 (NIE_3) is estimated to be very close to 0 (posterior mean (95% posterior interval) 0.02 (-0.04, 0.08)). Estimates of the joint indirect effects that involve SO_2 ($JNIE_{12}$ and $JNIE_{13}$) are close in magnitude to that of $JNIE_{123}$. The estimated $JNIE_{12}$ is nearly identical to the $JNIE_{123}$, indicating that reductions in both SO_2 and NO_x account for nearly all of the joint mediated effect of all emissions. The estimated $JNIE_{13}$ is nearly identical to NIE_1, indicating that combining causal reductions of SO_2 with causal reductions of CO_2 does not significantly change the mediated effect over what is due to reductions in SO_2 alone.

The overall conclusion of the causal mediation analysis is that 1) SO_2 scrubbers have a strong direct effect on ambient $PM_{2.5}$ that is not mediated through changes in emissions of SO_2 , NO_x , or CO_2 ; 2) slightly over one third of the total effect of scrubbers on $PM_{2.5}$ is mediated through changes in these emissions, with a a strong indirect effect due to reductions in SO_2 accounting for most of the indirect mediated effect; 3) there is some evidence of a mediated effect due to changes in NO_x emissions, and 4) there is no evidence of mediation due to reductions in CO_2 emissions. Appendix B.5 provides an examination of the extent to which the indirect effects of the three emissions overlap with one another.

Conclusion and Discussion of Case Study 2

In this case study, we have evaluated the effectiveness of a specific regulatory intervention (presence of an SO_2 scrubber on a coal-fired power plant in 2005) in terms of 1) the intervention's causal effect on annual emission of

Table 6: Posterior mean and 95% probability intervals for total, direct, and indirect effects of SO_2 scrubbers.

	TE	$JNIE_{123}$	NDE
Mean	-1.179	-0.444	-0.735
95% C.I.	(-2.091, -0.290)	(-0.697, -0.231)	(-1.666, 0.176)
	NIE_1	NIE_2	NIE_3
Mean	-0.364	-0.097	0.019
95% C.I.	(-0.530, -0.220)	(-0.257, 0.063)	(-0.038, 0.080)
	$JNIE_{12}$	$JNIE_{23}$	$JNIE_{13}$
Mean	-0.462	-0.079	-0.345
95% C.I.	(-0.706, -0.271)	(-0.242, 0.091)	(-0.509, -0.191)

SO_2 , NO_x , and CO_2 in 2005 2) the intervention's causal effect on average annual ambient $PM_{2.5}$ in 2005 among monitors within 150km of a power plant, and 3) the extent to which the intervention's causal effect on ambient $PM_{2.5}$ is mediated through causal reductions in multiple emissions. We focus in particular on the question of mediation to provide the first empirical evidence of the presumed causal relationships that motivate emissions control interventions that were a key part of the ARP and continue as important strategies for improving ambient air quality and, ultimately, human health. Given that our questions of interest - and indeed many accountability questions - pertain to mediated effects of multiple pollutants that are measured contemporaneously and possibly interact with one another, we developed new methods for principal stratification and causal mediation analysis for multiple contemporaneous and non-independent mediators. We introduced Bayesian nonparametric modeling and estimation techniques to provide flexible models to the observed data, and linked observed data distributions into joint distributions of potential outcomes using explicit and transparent assumptions (mostly contained in Appendix B.3).

A key feature of our case study is the integration of a principal stratification analysis and a causal mediation analysis that rely on the exact same modeling assumptions for the observed data. The difference between these two analysis pertains to the presence or absence of assumptions about potential outcomes that are *unobservable* for every member of the study sample. Thus, we begin with a principal stratification analysis relying on assumptions for the observable outcomes $Y(z,\mathbf{M}(z,z,z)),\mathbf{M}(z,z,z)$ to identify the principal causal effects, and then augment these assumptions with assumptions about unobservable potential outcomes (e.g., $Y(1,\mathbf{M}(0,0,0))$) to

estimate mediation effects. The explicit connection between principal stratification and causal mediation analyses explored here represents, to our knowledge, the most comprehensive consideration of these two approaches and the implications of their results in the context of a single analysis.

The results of the principal stratification and causal mediation analyses should be interpreted jointly, and are, in this case study, largely consistent with one another. The estimated dissociative effects were significantly less than zero for all combinations of emissions, indicating that scrubbers reduced $PM_{2.5}$ even without changing emissions significantly. This is consistent with the result of the mediation analysis where the estimated NDE accounted for nearly two thirds of the total effect of a scrubber on ambient $PM_{2.5}$. The difference between associative and dissociative effects was most pronounced when considering emissions of SO_2 , either individually or in combination with other emissions. This indicated that power plants that exhibited large causal effects on SO_2 emissions also exhibited large effects on ambient $PM_{2.5}$. This conclusion is consistent with the results of the mediation analysis where the NIE of SO_2 and the the JNIEs involving SO_2 ($JNIE_{12}$, $JNIE_{13}$, and $JNIE_{123}$) were all significantly less than zero and similar in magnitude to one another. Furthermore, the magnitude of the associative effects relative to NDE and NIEs are consistent with the well-known result that, in general, associative effects are a mixture of direct and indirect effects. For example, the EAE for plants that exhibited large reductions in SO_2 was estimated to be -1.402, which is similar to the sum of the estimates of NDE$= -0.74$ and $NIE_1 = -0.36$.

Interpretation of the results of this case study should be viewed in light of several important limitations. First is the relative simplicity with which we linked power plants to monitors. Specifically, our strategy simply links power plants to all of the ambient monitors within 150km. Thus, our analysis is of the causal effects of scrubbers on average $PM_{2.5}$ measured within 150km, which likely does not reflect the full effect of emissions changes from power plants on ambient air quality, which are expected to have implications at distances much greater than 150km. Furthermore, when a monitor is located within 150km of more than one power plant, it is assigned to its closest power plant, which results in many power plants that are assigned to no monitors being excluded from our analysis. Thus, our analysis assumes that impacts on ambient $PM_{2.5}$ due to emissions changes in these excluded power plants are distributed evenly across areas surround the scrubber and no scrubber power plants included in the analysis. More sophisticated strategies to link ambient monitors to power plants based on features such as

atmospheric conditions and weather patterns are warranted, but analysis of the data constructed here represents an important approximation that still yields valuable inferences, especially with respect to quantifying causal pathways.

Another important limitation of this analysis is that it assumes that the factors listed in Table 4 are sufficient to control for confounding, which in this case would consist of differences between power plants or other features related to ambient $PM_{2.5}$ that are also associated with whether a power plant had a scrubber installed in 2005.

To develop our new methods we also considered analysis of a single year and regarded a power plant as "treated" if it had a scrubber installed in January 2005, without regard to how long the scrubber had been installed or changes in emissions and ambient $PM_{2.5}$ over time. Future work will develop a framework to accommodate longitudinal analysis by using Bayesian dynamic models, which could update information from the past and smooth the effects over the course of a several year time period (Kim et al., 2015).

The results of our analysis are largely consistent with expectations: SO_2 scrubbers appear to causally reduce ambient $PM_{2.5}$ (within 150km) and this causal effect is mediated to a considerable extent by causal reductions in SO_2 emissions and not to considerably mediated through reductions in NO_x or CO_2 emissions. The finding that there appears to be a considerable direct effect of SO_2 scrubbers on $PM_{2.5}$ that is not mediated through emissions of SO_2 , NO_x , or CO_2 is somewhat surprising. More work is warranted to learn about possible explanations of this direct effect, which could relate to reduction of other emissions (e.g., primary particles), features of secondary formation of $PM_{2.5}$ that are not directly captured by direct emissions, or simply the attribution of direct effects to phenomena that truly relate to unmeasured confounding.

Despite the limitations of this case study, we have provided the first empirical investigation of the presumed causal pathways that motivate a variety of air quality control strategies that aim to reduce harmful emissions from power plants. Using a principled causal inference framework and rigorous analysis to quantify causal pathways, we have evaluated the effectiveness of scrubber installation for reducing emissions and ambient $PM_{2.5}$, representing an analysis of two important links in the chain of accountability amid the realities of actual regulatory implementation. The health implications of our analysis rely on the presumed link between ambient $PM_{2.5}$ and health outcomes, but the methods presented here can be applied in other multipollutant accountability settings,

including extensions of the current analysis to investigate, for example, the extent to reductions in multiplollutant emissions mediate causal health effects or the extent to which scrubber-induced changes in ambient $PM_{2.5}$ (or other pollutants) mediate causal effects on health outcomes.

Additional Work in Progress

In addition to the analysis of Case Study 2, we have a variety of ongoing analyses of causal impacts of power plant emissions controls. We have rigorously evaluated the causal impact of SO_2 and NO_x control strategies on SO_2 and NO_x emissions among 995 coal-burning EGUs during the years 1997-2012. We are also extending the analysis of Case Study 2 to investigate the extent to which the causal effect of SO_2 scrubber installation on Medicare health outcomes is mediated through reductions in ambient $PM_{2.5}$, which requires more sophisticated procedures for linking data between power plants, ambient monitors, and residential zip codes of Medicare beneficiaries.

Discussion and Conclusions

Over the past ten years, important progress in accountability assessment has initiated a new dimension to the scientific evidence available for informing policy decisions. While important challenges remain, the perspectives and methods in this report represent progress towards rigorous evaluation of large-scale regulatory policies. Sharpening the distinction between analytic perspectives for exposure-response estimation and for estimating causal effects of well-defined actions is necessary in order to advance accountability assessment beyond evaluation of localized, abrupt actions and towards informing policy debates with evidence of the effects of broad and complex regulations. We have outlined the particular relevance of potential outcomes methods for causal inference for advancing the goals of accountability assessment to focus on the direct evaluation of the effectiveness of specific policies or actions. Our analysis in the Case Study 1: Accountability Assessment of PM_{10} Nonattainment Designations Section illustrated how potential-outcomes reasoning can be deployed towards the goals of long-term direct accountability assessment. The Case Study 2: Accountability Assessment of Power Plant Emissions Controls Section outlined the development of new statistical methods for multipollutant accountability assessment, and illustrated how potential-outcomes perspectives can be useful for quantifying various causal pathways through which an air quality intervention impacts outcomes.

The deployment of potential outcomes methods for direct accountability assessment represents an important new direction for accountability research and air pollution epidemiology more broadly. Defining causal consequences of well-defined actions, what we refer to as direct accountability, stands in important contrast to the study of how air pollution exposure relates to the onset of clinical disease. Thus, the analytic perspectives and associated statistical methods here are consistent with a recent emphasis on so-called "consequentialist epidemiology" that turns focus away from identifying underlying causes of disease and towards development of consequential interventions (Galea, 2013). This is not to diminish the importance of a vast array of air pollution epidemiological evidence that motivates the need to intervene to control population exposures, but rather to emphasize the need to provide equally strong evidence of the consequences of specific interventional strategies to protect public health and the environment. While no single analytic strategy can overcome all the challenges inherent to accountability,

assessment the best science should be generated from a variety of available approaches. We argue that rigorous efforts to directly evaluate causal effects of well-defined regulatory interventions constitute one such approach that, while distinct from traditional epidemiological tools, is essential to the current regulatory climate.

References

Banzhaf, H. S., Burtraw, D., Evans, D., and Krupnick, A. (2006). Valuation of natural resource improvements in the Adirondacks. *Land Economics* **82,** 445–464.

Buonocore, J. J., Dong, X., Spengler, J. D., Fu, J. S., and Levy, J. I. (2014). Using the Community Multiscale Air Quality (CMAQ) model to estimate public health impacts of PM2.5 from individual power plants. *Environment International* **68,** 200–208.

Burtraw, D. (1999). Cost savings, market performance and economic benefits of the US Acid Rain Program. *Pollution for Sale. Emissions Trading and Joint Implementation, Cheltenham* pages 43–62.

Burtraw, D., Krupnick, A., Mansur, E., Austin, D., and Farrell, D. (1998). Costs and benefits of reducing air pollutants related to acid rain. *Contemporary Economic Policy* **16,** 379–400.

Cefalu, M. and Zigler, C. M. (2015). Posterior-predictive treatment assignments and the estimation of causal effects. *Technical Report* .

Chay, K., Dobkin, C., and Greenstone, M. (2003). The Clean Air Act of 1970 and Adult Mortality. *Journal of Risk and Uncertainty* **27,** 279–300.

Chen, Y., Ebenstein, A., Greenstone, M., and Li, H. (2013). Evidence on the impact of sustained exposure to air pollution on life expectancy from Chinas Huai River policy. *Proceedings of the National Academy of Sciences* **110,** 12936–12941.

Chestnut, L. G. and Mills, D. M. (2005). A fresh look at the benefits and costs of the US acid rain program. *Journal of Environmental Management* **77,** 252–266.

Clancy, L., Goodman, P., Sinclair, H., and Dockery, D. W. (2002). Effect of air-pollution control on death rates in Dublin, Ireland: an intervention study. *The Lancet* **360,** 1210–1214.

Correia, A. W., Pope, C. A., Dockery, D. W., Wang, Y., Ezzati, M., and Dominici, F. (2013). Effect of Air Pollution Control on Life Expectancy in the United States: An Analysis of 545 U.S. Counties for the Period from 2000 to 2007. *Epidemiology* **24,** 23–31.

Crosas, M. (2011). The Dataverse Network: An Open-Source Application for Sharing, Discovering and Preserving Data. *D-Lib Magazine* **17,**.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96,** 187 –199.

Currie, J. and Walker, R. (2011). Traffic Congestion and Infant Health: Evidence from E-ZPass. *American Economic Journal: Applied Economics* **3,** 65–90.

Daniel, R. M., De Stavola, B. L., Cousens, S. N., and Vansteelandt, S. (2014). Causal mediation analysis with multiple mediators. *Biometrics* .

Deschenes, O., Greenstone, M., and Shapiro, J. S. (2012). Defensive Investments and the Demand for Air Quality: Evidence from the NOx Budget Program and Ozone Reductions. Working Paper 18267, National Bureau of Economic Research, Cambridge, MA.

Dockery, D. W., Pope III, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris, B. G., and Speizer, F. E. (1993). An association between air pollution and mortality in six US cities. *New England journal of medicine* **329,** 1753–1759.

Dominici, F., Greenstone, M., and Sunstein, C. R. (2014). Particulate Matter Matters. *Science* **344,** 257–259.

Dominici, F., Peng, R. D., Bell, M. L., Pham, L., McDermott, A., Zeger, S. L., and Samet, J. M. (2006). Fine Particulate Air Pollution and Hospital Admission for Cardiovascular and Respiratory Diseases. *JAMA: The Journal of the American Medical Association* **295,** 1127 –1134.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association* **90,** 577–588.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58,** 21–29.

Friedman, M. S., Powell, K. E., Hutwagner, L., Graham, L. M., and Teague, W. G. (2001). Impact of changes in transportation and commuting behaviors during the 1996 Summer Olympic Games in Atlanta on air quality and childhood asthma. *JAMA: the journal of the American Medical Association* **285,** 897–905.

Galea, S. (2013). An Argument for a Consequentialist Epidemiology. *American Journal of Epidemiology* page kwt172.

Glass, T. A., Goodman, S. N., Hernn, M. A., and Samet, J. M. (2013). Causal inference in public health. *Annual review of public health* **34,** 61–75.

Greenstone, M. (2004). Did the Clean Air Act cause the remarkable decline in sulfur dioxide concentrations? *Journal of Environmental Economics and Management* **47,** 585–611.

Greenstone, M. and Gayer, T. (2009). Quasi-experimental and experimental approaches to environmental economics. *Journal of Environmental Economics and Management* **57,** 21–44.

Health Effects Institute (2010). *Proceedings of an HEI Workshop on Further Research to Assess the Health Impacts of Actions Taken to Improve Air Quality*. Health Effects Institute, Boston, MA.

Hedley, A. J., Wong, C.-M., Thach, T. Q., Ma, S., Lam, T.-H., and Anderson, H. R. (2002). Cardiorespiratory and all-cause mortality after restrictions on sulphur content of fuel in Hong Kong: an intervention study. *The Lancet* **360,** 1646–1652.

HEI Accountability Working Group (2003). *Assessing the Health Impact of Air Quality Regulations: Concepts and Methods for Accountability Research. Communication 11.* Health Effects Institute, Boston, MA.

Hernan, M. A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Willett, W. C., Manson, J. E., and Robins, J. M. (2008). Observational Studies Analyzed Like Randomized Experiments. *Epidemiology* **19,** 766–779.

Hidy, G., Brook, J., Demerjian, K., Pennell, W., and Scheffe, R. (2011). *Technical Challenges of Multipollutant Air Quality Management*. Springer, New York, NY.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis* **15,** 199 –236.

Hubbell, B. (2012). Assessing the Results of Air Quality Management Programs. *EM Magazine* pages 8–15.

Imai, K. and Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis* **21,** 141–171.

Jara, A., Hanson, T. E., Quintana, F. A., Mller, P., and Rosner, G. L. (2011). DPpackage: Bayesian Nonand Semiparametric Modeling in R.

Joffe, M. M. and Greene, T. (2009). Related Causal Frameworks for Surrogate Outcomes. *Biometrics* **65,** 530–538.

Kim, C., Daniels, M. J., Roy, J. A., and Marcus, B. H. (2015). Longitudinal Causal Mediation Analysis for Behaviorial Trials. *Technical Report* .

King, G. (2007). An introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods & Research* **36,** 173–199.

King, G. and Zeng, L. (2006). The Dangers of Extreme Counterfactuals. *Political Analysis* **14,** 131 –159.

Krewski, D., Burnett, R., Goldberg, M., Hoover, B. K., Siemiatycki, J., Jerrett, M., Abrahamowicz, M., and White, W. (2003). Overview of the reanalysis of the Harvard six cities study and American Cancer Society study of particulate air pollution and mortality. *Journal of Toxicology and Environmental Health Part A* **66,** 1507–1552.

Laden, F., Schwartz, J., Speizer, F. E., and Dockery, D. W. (2006). Reduction in Fine Particulate Air Pollution and Mortality: Extended follow-up of the Harvard Six Cities Study. *Am. J. Respir. Crit. Care Med.* pages 200503–443OC.

Levy, J. I., Baxter, L. K., and Schwartz, J. (2009). Uncertainty and Variability in Health-Related Damages from Coal-Fired Power Plants in the United States. *Risk Analysis* **29,** 1000–1014.

Levy, J. I., Wilson, A. M., and Zwack, L. M. (2007). Quantifying the efficiency and equity implications of power plant air pollution control strategies in the United States. *Environmental health perspectives* pages 743–750.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Routledge.

Moore, K., Neugebauer, R., Lurmann, F., Hall, J., Brajer, V., Alcorn, S., and Tager, I. (2010). Ambient Ozone Concentrations and Cardiac Mortality in Southern California 19832000: Application of a New Marginal Structural Model Approach. *American Journal of Epidemiology* **171,** 1233 –1243.

Mller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83,** 67–79.

National Research Council (2004). *Air Quality Management in the United States*. National Academies Press, Washington, DC.

Nelsen, R. B. (1999). *An introduction to copulas*, volume 139. Springer Science & Business Media.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pages 411–420.

Pearl, J. (2011). Principal stratificationa goal or a tool? *The International Journal of Biostatistics* **7,**.

Peng, R. D., Chang, H. H., Bell, M. L., McDermott, A., Zeger, S. L., Samet, J. M., and Dominici, F. (2008). Coarse Particulate Matter Air Pollution and Hospital Admissions for Cardiovascular and Respiratory Diseases Among Medicare Patients. *JAMA: The Journal of the American Medical Association* **299,** 2172 –2179.

Pope III, C. A. (1996). Particulate pollution and health: a review of the Utah valley experience. *Journal of exposure analysis and environmental epidemiology* **6,** 23–34.

Pope III, C. A., Ezzati, M., and Dockery, D. W. (2009). Fine-particulate air pollution and life expectancy in the United States. *New England Journal of Medicine* **360,** 376–386.

Pope III, C. A., Rodermund, D. L., and Gee, M. M. (2007). Mortality Effects of a Copper Smelter Strike and Reduced Ambient Sulfate Particulate Matter Air Pollution. *Environmental Health Perspectives* **115,** 679–683.

Rich, D. Q., Kipen, H. M., Huang, W., Wang, G., Wang, Y., Zhu, P., Ohman-Strickland, P., Hu, M., Philipp, C., and Diehl, S. R. (2012). Association between changes in air pollution levels during the Beijing Olympics and biomarkers of inflammation and thrombosis in healthy young adults. *JAMA* **307,** 2068–2078.

Robins, J. M. and Greenland, S. (1992). Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology* **3,** 143–155.

Robins, J. M., Hernan, M., and Brumback, B. (2000). Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology* **11,** 550.

Rosenbaum, P. R. (1984). The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment. *Journal of the Royal Statistical Society. Series A (General)* **147,** 656–666.

Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **70,** 41–55.

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

Rubin, D. B. (2004). Direct and Indirect Causal Effects via Potential Outcomes. *Scandinavian Journal of Statistics* **31,** 161–170.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics* **2,** 808–840.

Samet, J. M. (2011). The Clean Air Act and Health A Clearer View from 2011. *New England Journal of Medicine* **365,** 198–201.

Schmalensee, R. and Stavins, R. (2012). The SO2 allowance trading system: The ironic history of a grand policy experiment. Technical report, National Bureau of Economic Research.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science : a review journal of the Institute of Mathematical Statistics* **25,** 1–21.

Tonne, C., Beevers, S., Armstrong, B., Kelly, F., and Wilkinson, P. (2008). Air pollution and mortality benefits of the London Congestion Charge: spatial and socioeconomic inequalities. *Occupational and Environmental Medicine* **65,** 620 –627.

U.S. EPA (2009). *Integrated Science Assessment for Particulate Matter*. U.S. Environmental Protection Agency Office of Research and Development, Research Triangel Park, NC.

U.S. EPA (2010). The Benefits and Costs of the Clean Air Act: 1990 to 2020. Revised Draft Report, US EPA Office of Air and Radiation, Research Triangel Park, NC.

U.S. EPA (2012). Environmental Benefits Mapping and Analysis Program (BenMAP) User's Manual. Technical report, US EPA.

U.S. EPA (2013). Workshop on Designing Research to Assess Air Quality and Health Outcomes from Air Pollution Regulations. In *Designing Research to Assess Air QuAlity and Health Outcomes from Air Pollution Regulations*, Research Triangel Park, NC.

van Erp, A. M., Cohen, A. J., Shaikh, R., and O'Keefe, R. (2012). Recent progress and challenges in assessing the effectiveness of air quality interventions towards improving public health: the HEI experience. *EM Magazine* pages 22–28.

VanderWeele, T. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface* **2,** 457–468.

VanderWeele, T. and Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic methods* **2,** 95–115.

VanderWeele, T. J. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics & Probability Letters* **78,** 2957–2962.

VanderWeele, T. J. (2009). Marginal Structural Models for the Estimation of Direct and Indirect Effects. *Epidemiology* **20,** 18–26.

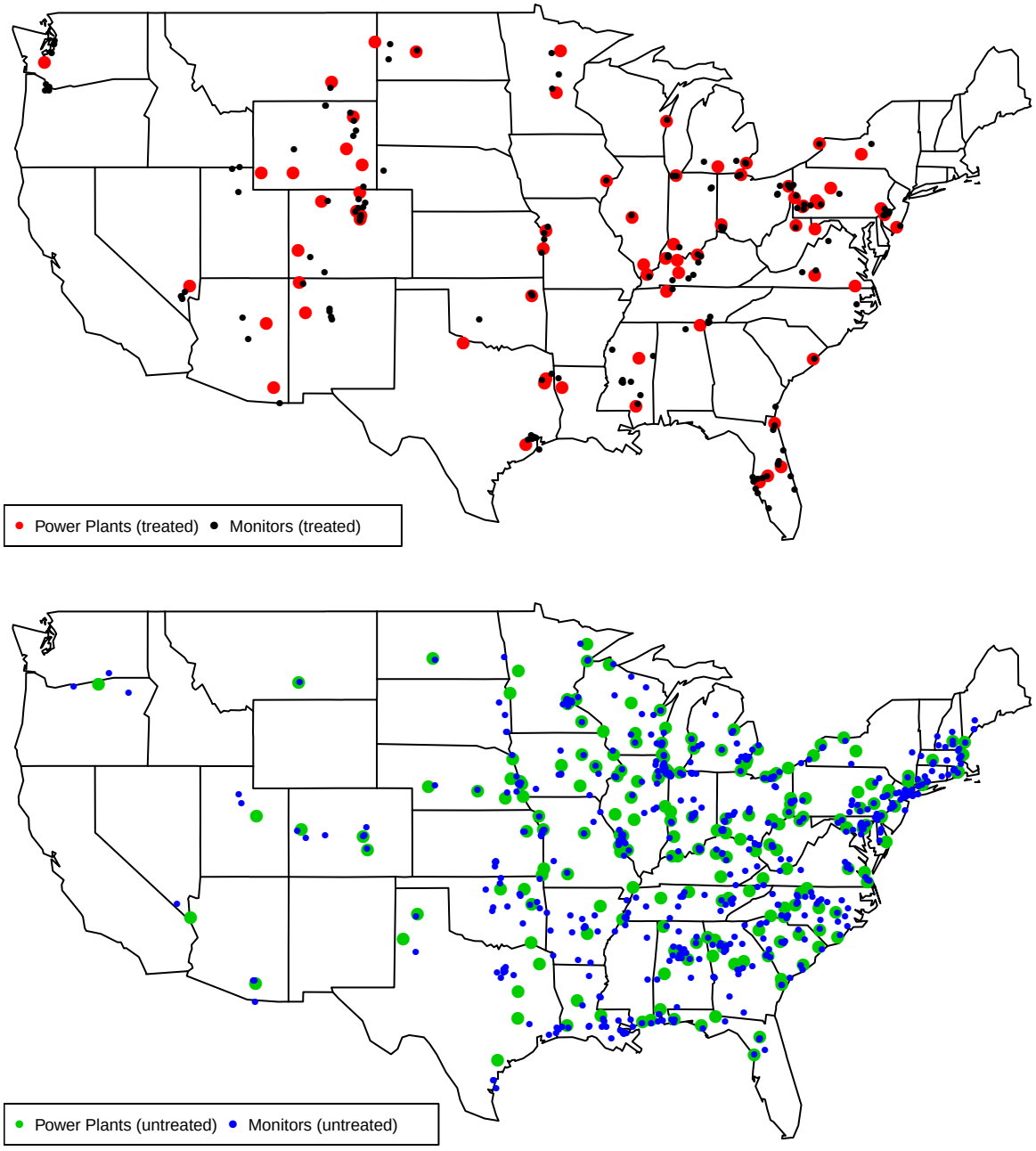
Zanobetti, A., Dominici, F., Wang, Y., and Schwartz, J. D. (2014). A national case-crossover analysis of the short-term effect of PM2. 5 on hospitalizations and mortality in subjects with diabetes and neurological disorders. *Environmental Health* **13,** 38.

Zeger, S. L., Dominici, F., McDermott, A., and Samet, J. M. (2008). Mortality in the Medicare Population and Chronic Exposure to Fine Particulate Air Pollution in Urban Centers (2000-2005). *Environmental Health Perspectives* **116,** 1614–1619.

Zigler, C. M. and Dominici, F. (2014). Point: clarifying policy evidence with potential-outcomes thinking – beyond exposure-response estimation in air pollution epidemiology. *American journal of epidemiology* **180,** 1133–1140.

Zigler, C. M., Dominici, F., and Wang, Y. (2012). Estimating causal effects of air quality regulations using principal stratification for spatially correlated multivariate intermediate outcomes. *Biostatistics (Oxford, England)* **13,** 289–302.

Figure 9: Locations of power plants and linked ambient $PM_{2.5}$ monitors for the analysis of SO_2 scrubber effects on emissions and ambient $PM_{2.5}$.



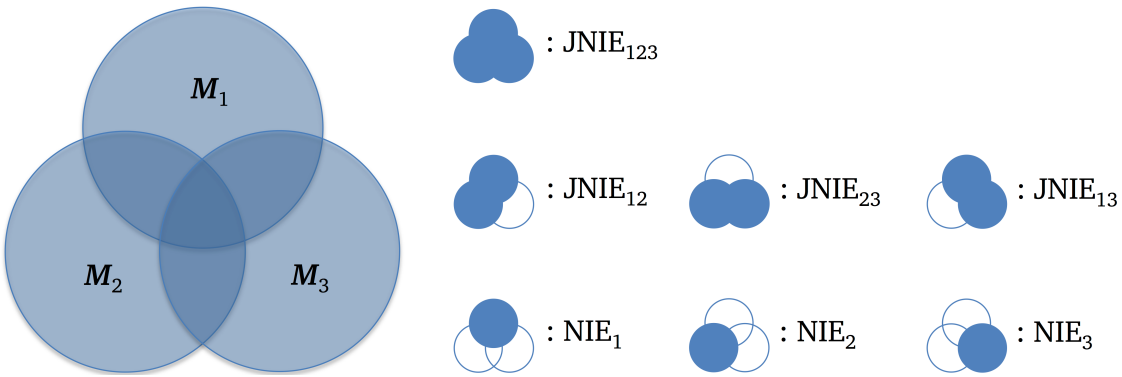


Figure 10: Graphical representation of partitioning the $JNIE_{123}$ for 3 mediators

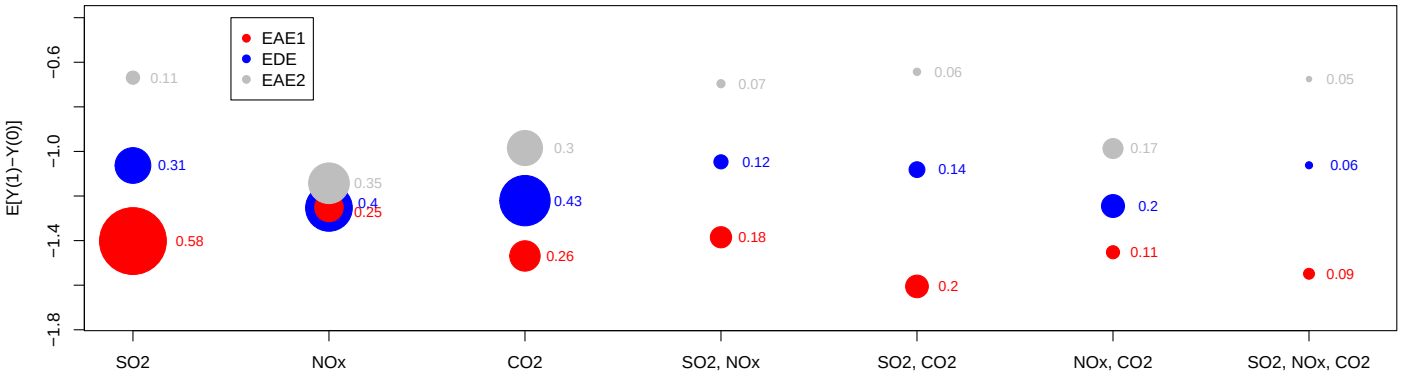
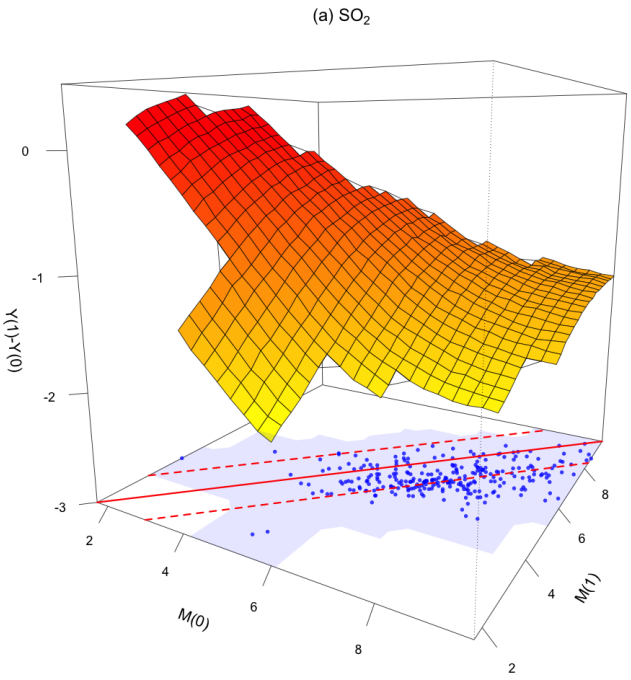
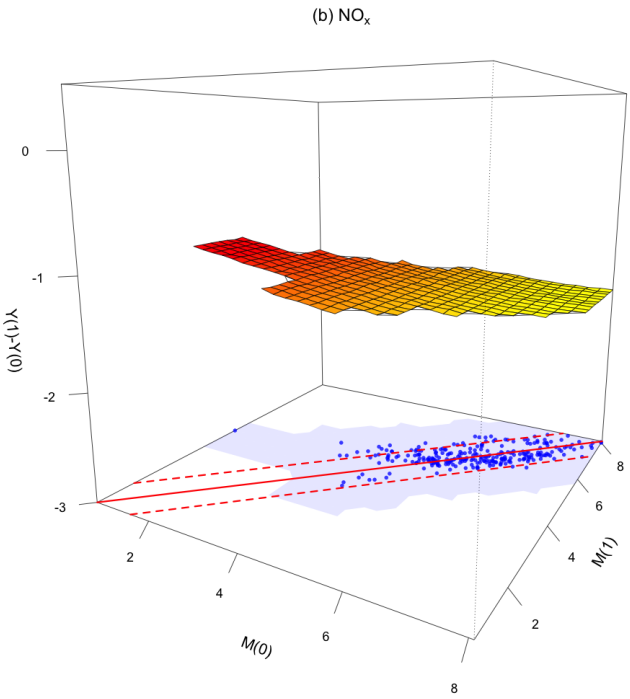


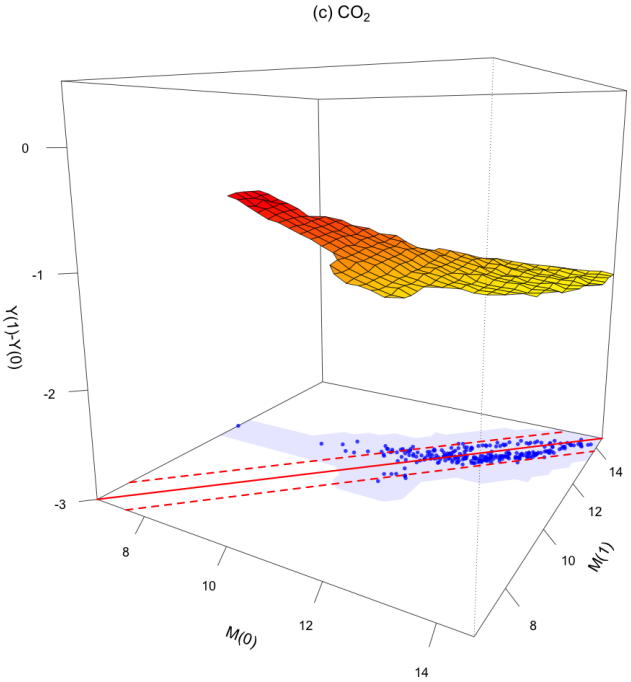
Figure 11: Posterior mean estimates of average associative and dissociative effects of SO_2 scrubbers. Size of circle is proportional to the percent of observations estimated to belong in the corresponding strata, and number listed is posterior mean proportion.



(a) $k = 1$ (SO_2)



(b) $k = 2$ (NO_x)



(c) $k = 3$ (CO_2)

Figure 12: Average surface plots of the causal effect on $PM_{2.5}$ for different values of $(M_k(0), M_k(1))$. Values of $(M_k(0), M_k(1))$ are plotted on the x- and y- axes, and determine the causal effect of a scrubber on emission k. The corresponding value of the causal effect of a scrubber on $PM_{2.5}$ $(Y(1) - Y(0))$ is plotted on the z-axis. The blue cloud of points are simulations of $(M_k(0), M_k(1))$ for one MCMC iteration to represent the range of values of $(M_k(0), M_k(1))$ consistent with the observed data. Red lines are at $M_k(0) = M_k(1)$ (solid line) and $+/-\sigma_k$ (dashed lines).

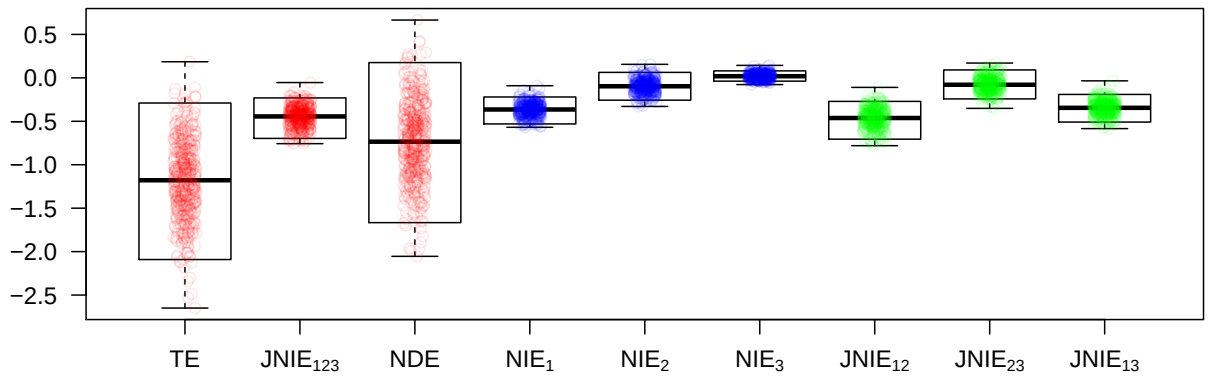


Figure 13: Posterior distributions of direct and indirect effects in the analysis of SO_2 scrubbers.