Why Acute Health Effects of Air Pollution Could Be Inflated [†]

Vincent Bagilet[1]

Léo Zabrocki[2]

September 23, 2021

Abstract

Accurately measuring the short-term effects of air pollution on health plays a key role in setting air quality standards. Yet, statistical power calculations are rarely—if ever—carried out. We first collect estimates and standard errors of all available articles found in the standard epidemiology and causal inference literatures. We find that nearly half of them may suffer from a low statistical power and could thereby produce statistically significant estimates that are actually inflated. We then run simulations based on real data to identify which parameters of research designs affect statistical power. Despite their large sample sizes, we show that studies exploiting rare exogenous shocks such as transport strikes or thermal inversions could have a very low statistical power, even for plausibly large effect sizes. Our simulation results indicate that the observed discrepancy in the literature between instrumental variable estimates and non-causal ones could be partly explained by the inherent imprecision of the two-stage least-squares estimator. We also provide evidence that subgroup analysis on the elderly or children should be implemented with caution since the average number of events for an health outcome is a major driver of power. Based on these findings, we build a series of recommendations for researchers to evaluate the design of their study with respect to statistical power issues.

[†]**Comments and suggestions are highly welcome.** We are very grateful to Hélène Ollivier and Jeffrey Shrader for their fantastic guidance on this project. Many thanks to Geoffrey Barrows, Marie-Abèle Bind, Sylvain Chabé-Ferret, Marion Leroutier, Quentin Lippmann, as well as seminars participants at the SusDev Colloquium at Columbia, IPWSD, M&A's Lab, FAERE for their feedbacks.

[1]Columbia University - SIPA, New York, US. Email: vincent.bagilet@columbia.edu

[2]Paris School of Economics and École des Hautes Etudes en Sciences Sociales, Paris, France. Email: leo.zabrocki@psemail.eu

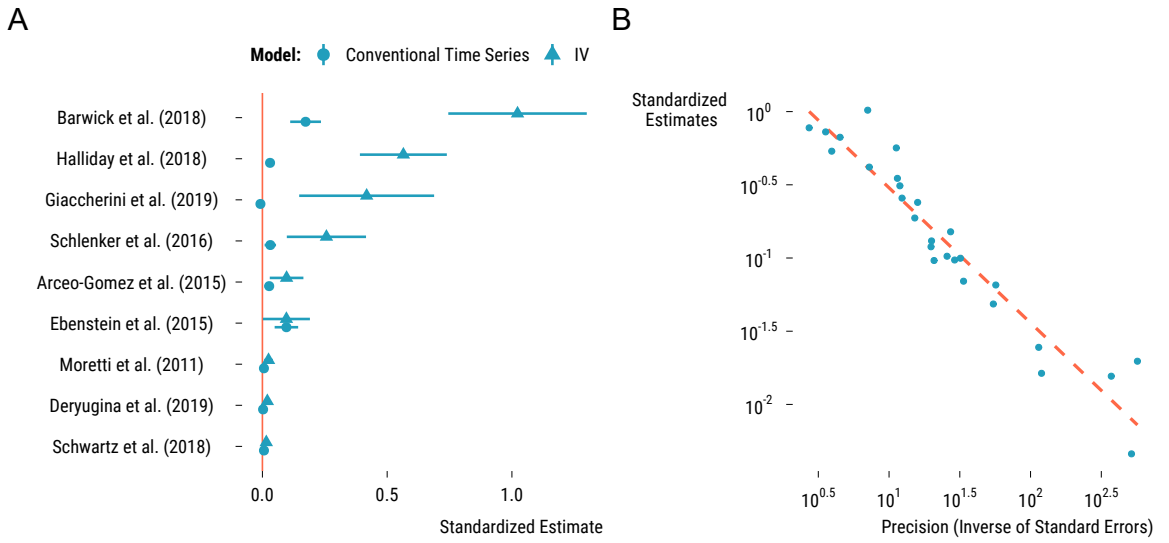
1 Introduction

From extreme events such as the London Fog of 1952 to the development of sophisticated time-series analyses, a vast scientific literature in epidemiology has established that air pollution induces adverse health effects on the very short-term (Schwartz 1994, Le Tertre et al. 2002, Bell et al. 2004, Di et al. 2017, Liu et al. 2019). Increases in the concentration of several ambient air pollutants have been found to be associated with small relative increases in the daily mortality and emergency admissions for respiratory and cardiovascular causes (Samet et al. 2000, Shah et al. 2015, Orellano et al. 2020). All this evidence led to the implementation of public policies such as air quality alerts to mitigate the acute effects of air pollutants. Accurate estimates of the short-term health effects of air pollution are therefore crucial as they directly inform public health policies.

With this objective in mind, researchers in economics and epidemiology have addressed the issue of unmeasured confounding variables with causal inference methods in the last decade (Dominici and Zigler 2017, Bind 2019). Newly obtained results confirm the acute health effects of air pollution (Schwartz et al. 2015; 2018, Deryugina et al. 2019). Yet, causal estimates are often larger than what would have been predicted by the standard epidemiology literature. For instance, in Panel A of Figure 1, we see that instrumental variable estimates of 9 studies in the causal inference literature based on this method are always larger than naive estimates. This could arguably be explained by the fact that instrumental strategies remove omitted variable bias and reduce attenuation bias coming from classical measurement error in air pollution exposure. Panel B of Figure 1 however suggests an alternative explanation. For the 29 papers using causal inference methods found in this literature, we plot the standardized estimates against the inverse of their standard errors, which is a proxy for a study's precision. Large effect sizes are only found in

imprecise studies and the more precise the study, the smaller the effect size. The negative relationship between effect sizes and studies' precision has also been observed in fields such as medicine, psychology and economics (Button et al. 2013, Camerer et al. 2018, Schäfer and Schwarz 2019).

Figure 1: Naive versus Causal Estimates and the Deflation of Effect Sizes as Precision Increases.



Notes: In Panel A, standardized estimates and their associated 95% confidence intervals are displayed for the 9 articles of the causal inference literature based on instrumental variable strategies and for which estimates from naive regressions are available. Triangles represent instrumental variable estimates with dots are naive regression estimates. In panel B, standardized estimates of the 29 articles of the causal inference literature are plotted against the inverse of the standard errors, which can been considered as a measure of precision. Both axes are on a log10 scale.

Following Ioannidis (2008b) and Gelman and Carlin (2014), the variation in studies' statistical power could explain the origin of this negative relationship but also help understand why causal estimates are larger than those found in the epidemiology literature. Simply put, studies with low precision result in larger effect sizes. Their statistical power is low and, to be statistically significant, their estimates need to be large enough, at least 2 standard errors away from 0 at the 5% significance level. Since statistically significant results are more likely to be published, some estimates found in the literature may be inflated as they would come

from a non-representative sample of the estimates, those large enough to be statistically significant (Brodeur et al. 2016; 2020). The consequences of low statistical power are not specific to studies on short-term health effects of air pollution but may be particularly salient in this literature where the signal-to-noise ratio is often low (Peng et al. 2006, Peng and Dominici 2008).

In this paper, we undertake the first empirical investigation to determine if studies on the short-term health effects of air pollution could be under-powered and thereby produce inflated estimates. We start tackling this question by gathering a unique corpus of about 600 studies based on associations and 29 articles that rely on causal inference. For each of these papers, we run statistical power calculations to assess whether the design of the study would be robust enough to confidently detect an effect size smaller than the observed estimate (Gelman and Carlin 2014, Ioannidis et al. 2017, Liu et al. 2017, Timm 2019). Using real data from the US National Morbidity, Mortality, and Air Pollution Study (Samet et al. 2000), we then implement simulations to identify the characteristics of research designs that drive their statistical power and the inflation of statistically significant estimates (Gelman et al. 2020, Altoè et al. 2020).

The results of our statistical power calculations show that research designs based on associations and causal inference methods are similarly prone to statistical power issues. Half of the studies in the two strands of the literature have a statistical power below 80% to detect effect sizes 25% smaller than their observed estimates. Underpowered studies could produce statistically significant estimates respectively 1.4 and 2 times larger than true effect sizes in the causal inference and standard epidemiology literatures. Our retrospective power calculations also highlight a wide heterogeneity in the robustness of articles with respect to statistical power issues. For example, if the true effect sizes are equal to the ones predicted by the standard epidemiology literature, the statistical power of studies using instrumental variable

designs would range from 5% to 64%. In some studies, statistically significant estimates would be just 1.3 times larger than the true effect sizes, while in others, the inflation factor could be as high as 41.

Our simulation results help understand why some research designs face statistical power issues. We first show that a very large number of observations is needed for all causal inference methods to reach a sufficient statistical power. Regression discontinuity designs based on air quality alerts rely on sample sizes that are too small for statistically significant estimates not to be inflated. Second, we show that the use of public transport strikes or thermal inversions as exogenous shocks on air pollution could be problematic. These studies are based on rare events, which in some cases represent less than 1% of the observations. The resulting statistical power is very low, around 15%, and statistically significant estimates can exaggerate even large true effect sizes by a factor of 2.7. Third, we find that the average daily count of cases of an health outcome is a key driver of statistical power for all empirical strategies. Statistical significant estimates of the effects of air pollution on the elderly or children can be very inflated since health outcomes for these groups often have few daily cases.

Our article makes two contributions to the literature on the acute health effects of air pollution. First, as highlighted by the replication crises in medicine, psychology and experimental economics (Button et al. 2013, Collaboration et al. 2015, Camerer et al. 2018), there is a crucial need to evaluate the deficiencies of current statistical practices grounded in the null hypothesis significance testing framework (Ziliak and McCloskey 2008b, Simonsohn et al. 2014, Smaldino and McElreath 2016, Greenland 2017, Christensen et al. 2019, Amrhein et al. 2019). Our paper participates in the growing literature that uses retrospective power calculations to evaluate the plausibility of published findings (Ioannidis 2008a, Gelman and Carlin 2014, Smaldino and McElreath 2016, Ioannidis et al. 2017, Ferraro and Shukla

2020, Stommes et al. 2021). To the best of our knowledge, this paper is the first to show how to carry out statistical power calculations for studies on air pollution and human health. We also provide the first evidence that under-powered studies are an actual issue in this field.

Second, except for standard models used in the epidemiology literature (Winquist et al. 2012), few statistical power analyses exist to help researchers improve their studies design (Bhaskaran et al. 2013). This paper is the first to give concrete recommendations to avoid statistical power issues for several research designs estimating the acute health effects of air pollution. Statisticians have long advocated the use of fake-data simulation to flexibly evaluate the inference properties of statistical models (Gelman and Carlin 2014, Vasishth and Gelman 2019, Altoè et al. 2020, Gelman et al. 2020). In our paper, we follow this advice but rely instead on real-data since it is very complex to correctly simulate the relationships between ambient air pollution, weather parameters, calendar indicators and health outcomes. Our article is more closely connected to three recent articles evaluating the type I error rate and the lack of statistical power of several panel data models used to estimate the impacts of public policies on mortality outcomes (Schell et al. 2018, Black et al. 2019, Griffin et al. 2020). Treatment effects are on a longer time scale in these simulations but could be very useful to assess statistical power issues in studies on low-emission and congestion pricing zones. On the contrary, our simulations gauge the capacity of reduced-form, instrumental variable and regression discontinuity designs to estimate very short-run effects in the context of high-frequency data.

Finally, our contributions would not have any serious practical application if researchers cannot follow them. We therefore strive to make our analyses fully and easily reproducible to help them implement retrospective power calculations and power simulations in their own studies. We use state-of-the-art literate programming to explain and render all coding procedures in nicely formatted HTML

documents (Allaire et al. 2018). All replication and supplementary materials are available on this website.

In the following section, we implement a simple simulation exercise to show why statistically significant estimates exaggerate true effect sizes when studies have a low statistical power. In section 3, we present our retrospective analysis of the literature. In Section 4, we detail our simulation procedure to replicate empirical strategies. We display the simulation results in section 5 and provide specific guidance on study design in Section 6.

2 Background on Statistical Power, Type M and S errors

In a seminal paper, Gelman and Carlin (2014) point out that researchers working in the null hypothesis significance testing framework are often unaware that "statistically significant" estimates suffer from a winner's curse in under-powered studies: these estimates can largely overestimate true effect sizes and can even be of the opposite sign. In this section, we implement a simple simulation exercise to illustrate these two counter-intuitive issues and explain why they could matter in studies on the acute health effects of ambient air pollutants.

2.1 A Fictional Example

Imagine that a mad scientist is able to implement a randomized experiment to measure the short-term effects of air pollution on daily non-accidental mortality. The experiment takes place in a major Western city over the 366 days of a leap year. The scientist is able to increase concentration of particulate matter with a diameter below 2.5 µm ($PM_{2.5}$) by 10 µg/m^3—a large shock equivalent to one standard

deviation increase in the concentration of $PM_{2.5}$. Concretely, the scientist implements a complete experiment where they randomly allocate half of the days to the treatment group and the other half to the control group. They then measure the treatment effect of the intervention by computing the average difference in means between treated and control outcomes: the estimate for the treatment effect is equal to 4 additional deaths and is "statistically significant" at the 5% level, with a p-value of 0.04. The statistical significance of the estimate fulfills the scientist expectations, who immediately starts writing their paper. Had they not obtained a statistically significant estimate, they might not have submitted their result.

Unfortunately for the scientist, we know what the true effect of the experiment is since we created the data. In Table 1, we display the Science table where we observe the pair of potential outcomes of each day, $Y_i(W_i = 0)$ and $Y_i(W_i = 1)$. Y_i represents a daily count of non-accidental death and W_i the treatment assignment, which is equal to 1 for treated units and 0 otherwise. We first simulated the daily non-accidental mortality counts in the absence of treatment by drawing 366 observations from a negative binomial distribution with a mean of 106 and a variance of 402. We chose the parameters to approximate the distribution of non-accidental mortality counts in a large Western city. We then defined the counterfactual distribution of mortality by adding, on average, 1 extra death.

This treatment effect size represents approximately a 1% increase in the mean of the outcome. Note that the magnitude of this hypothetical effect is higher than what has been found in a recent and large-scale study based on 625 cities. Liu et al. (2019) found that a 10 $\mu g/m^3$ increase in $PM_{2.5}$ concentration was associated with a 0.68% (95% CI, 0.59 to 0.77) relative increase in daily all-causes mortality. Following the fundamental problem of causal inference, the daily count of deaths the scientist observe is given by the equation: $Y_i^{obs} = W_i \times Y_i(1) + (1 - W_i) \times Y_i(0)$. Treated units express their $Y_i(1)$ values and control units their $Y_i(0)$ values.

Table 1: Science Table of the Experiment.

Day Index	$Y_i(0)$	$Y_i(1)$	τ_i	W_i	Y_i^{obs}
1	122	124	+2	1	124
2	94	96	+2	1	96
3	96	98	+2	0	96
⋮	⋮	⋮	⋮	⋮	
364	96	97	+1	0	96
365	98	98	+0	0	98
366	143	144	+1	1	144

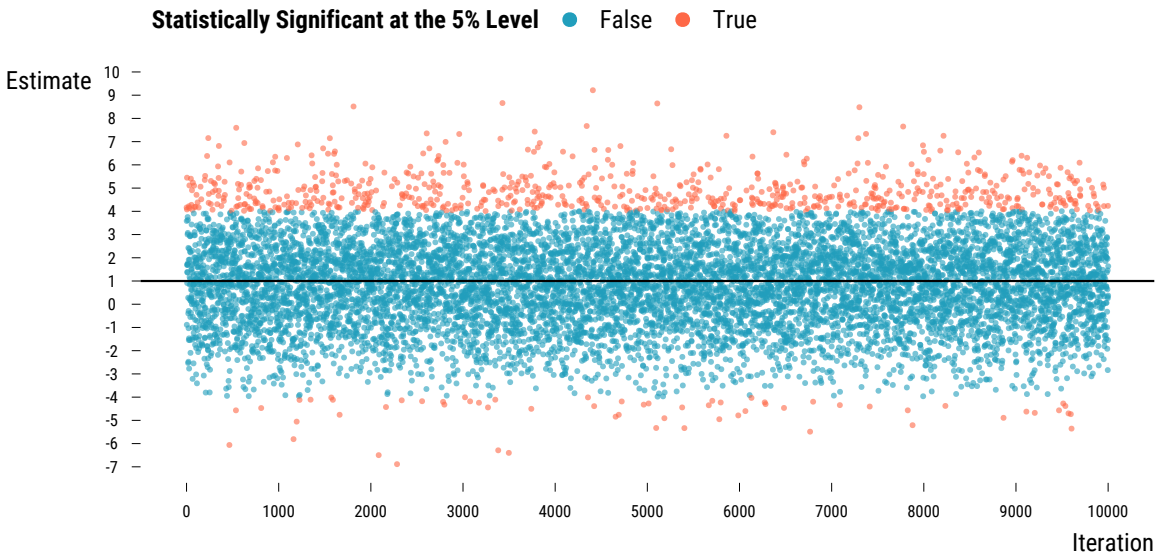
Notes: This table displays the potential outcomes, the unit-level treatment effect, the treatment status and the observed outcomes for 6 of the 366 daily units in the scientist's experiment.

With a random assignment of the treatment, how come the statistically significant estimate found by the scientist can be 4 times larger than the true treatment effect size? Replicating many times the experiment can help understand why.

2.2 Defining Statistical Power, Type M and S errors

In Figure 2, we plot the estimates of 10,000 iterations of the experiment. If there is a large variation in the effect size of estimates, the average is reassuringly equal to the true treatment effect of 1 additional death. We can however see that estimates close to the true effect size would not be statistically significant at the 5% level. In a world without publication bias, we could be confident that several replications of this experiment would recover the true treatment effect. Unfortunately, researchers are—despite recent changes in scientific practices and editorial policies—not incited enough to publish replication exercises and not statistically significant estimates. In a world with publication bias, only statistically significant estimates would be made public. Out of the 10,000 estimates, about 800 are statistically significant at the 5% level. The *statistical power* of the experiment, which can be defined as the probability to reject the null hypothesis against an alternative hypothesis, is

Figure 2: Estimates of the 10,000 Simulations.



Notes: In Panel A, blue and red dots represent the point estimates of the 10,000 iterations of the randomized experiment ran by the mad scientist. Red dots are statistically significant at the 5% level while blue dots are not. The black solid line represents the true average effect of 1 additional death.

therefore equal to 8%. The scientist was therefore very lucky to get a statistically significant estimate.

But with such a low statistical power, statistically estimates cannot be trusted anymore. Two metrics, the average type M (for magnitude) error and the probability to make a type S (for sign) error are useful to assess the negative consequences of lacking statistical power. First, we can evaluate by how much statistically significant estimates are inflated compared to the true treatment effect size by computing the average ratio of the absolute values of the statistically significant estimates over the true effect size (Gelman and Carlin 2014). With a statistical power of 8%, the scientist would on average make a type M error equal to 5! Second, we can notice that a non-negligible fraction of statistically significant estimates are of the wrong sign in Figure 2: this proportion is the probability of making a type S error (Gelman and Tuerlinckx 2000). For this experiment, a statistically significant estimate has a 8% probability of being of the wrong sign!

Thus, if the scientist would like to estimate the effect of the experiment through the prism of the statistical significance, they would need a larger number of observations: statistical power would then rise and conversely type M and S error would shrink.

2.3 Relevance for Studies on Acute Health Effects of Air Pollution

Type M and S errors are two concepts that highlight the danger of having too much confidence in statistically significant estimates when studies are under-powered. This issue is virtually absent from the literature on the acute health effect of air pollution but should be relevant for several reasons. First, researchers work with observational data and can often not control the sample size of their studies due to data availability. Very few guidance on the drivers of studies' statistical power actually exists (Winquist et al. 2012, Bhaskaran et al. 2013). Moreover, reaching a large statistical power could be challenging since estimated effect sizes are remarkably small and the modeling of high-frequency variations in daily mortality or emergency admission is difficult (Peng et al. 2006, Peng and Dominici 2008). Finally, we observe both in the standard epidemiology and the causal inference literatures a negative relationship between estimated effect sizes and studies' precision. It is important to investigate if this pattern could be explained by imprecise studies making type M errors (Ioannidis 2008b, Gelman and Carlin 2014, Ioannidis et al. 2017, Ferraro and Shukla 2020).

3 Retrospective Analysis of the Literature

In this section, we first explain how to implement a retrospective analysis of a study. Using different scenarios about the true effect sizes of studies found in the standard epidemiology and causal inference literatures, we then assess to which extent they

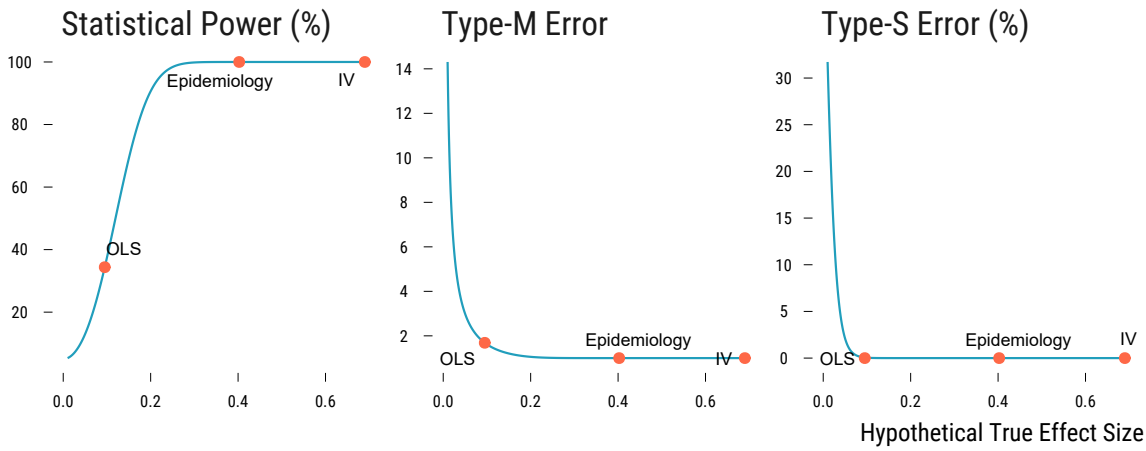
could suffer from low statistical power issues.

3.1 How to Run a Retrospective Analysis

Running a retrospective analysis only requires three metrics: the estimated effect, its standard error and a guess about the true effect size of the treatment of interest. Other parameters of the research design, such as the number of observations, are assumed to be fixed. Using the closed-form expressions derived by Liu et al. (2017) and their implementation in the R package `retrodesign` developed by Timm (2019), we can then compute the statistical power, the average exaggeration factor of statistically significant estimates and the probability that they are of the wrong sign. The usefulness of this analysis however relies entirely on a credible guess of the true effect size a study is trying to estimate. As the true effect is never observed, researchers can have very different priors on its magnitude. They could therefore assess differently the extent to which a study risks to suffer from statistical power issues. To illustrate this tension, we provide below a case study showing how a scientific discussion about effect sizes arises with a retrospective analysis.

In a flagship publication, Deryugina et al. (2019) instrument $PM_{2.5}$ concentrations with wind directions to estimate its effect on mortality, health care use, and medical costs among the US elderly. They gathered 1,980,549 daily observations at the county-level over the 1999–2013 period; it is one of the biggest sample sizes in the literature. When the authors instrument $PM_{2.5}$ with wind direction, they find that "a 1 $\mu g/m^3$ (about 10 percent of the mean) increase in $PM_{2.5}$ exposure for one day causes 0.69 additional deaths per million elderly individuals over the three-day window that spans the day of the increase and the following two days". The estimate's standard error is equal to 0.061. In Figure 3, we plot the statistical power, the inflation factor of statistically significant estimates and the probability that they are of the wrong sign as a function of hypothetical true effect sizes.

Figure 3: Power, Type M and S Errors Curves for Deryugina et al. (2019).



Notes: In each panel, a metric, such as the statistical power, the exaggeration ratio or the probability to make a type S error, is plotted against the range of hypothetical effect sizes. The "IV" label represents the value of the corresponding metric for an effect size equal to Deryugina et al. (2019)'s two-stage least square estimate. The "Epidemiology" label stands for the estimate found in Di et al. (2017), which is the epidemiology article most similar to Deryugina et al. (2019). The "OLS" label corresponds to the estimate found by Deryugina et al. (2019) when the air pollutant is not instrumented.

The estimate found by Deryugina et al. (2019) represents a relative increase of 0.18% in mortality. We labeled it as "IV" in Figure 3. Is this estimated effect size large compared to those reported in the standard epidemiology literature? We found a similar article to draw a comparison. Using a case-crossover design and conditional logistic regression, Di et al. (2017) find that a 1 $\mu g/m^3$ increase in $PM_{2.5}$ is associated with a 0.105% relative increase in all-cause mortality in the Medicare population from 2000 to 2012. The effect size found by Deryugina et al. (2019) is a bit higher than this estimate labeled as "Epidemiology" in Figure 3. If the estimate found by Di et al. (2017) was actually the true effect size of $PM_{2.5}$ on elderly mortality, the study of Deryugina et al. (2019) would have enough statistical power to perfectly avoid type M and S errors. Now, suppose that the true effect of the increase in $PM_{2.5}$ was 0.095 additional deaths per million elderly individuals—the estimate the authors found with a "naive" multivariate regression model. The statistical power would be 34%, the probability to make a type S error is null but the overestima-

tion factor would be on average equal to 1.7. Even with a sample size of nearly 2 million observations, Deryugina et al. (2019) could make a non-negligible type M error if the true effect size was the standard estimate. Yet, the authors could argue that their instrumental variable strategy leads to a higher effect size as it overcomes unmeasured counfounding bias. Besides, for effect sizes down to 0.182 additional deaths per million elderly individuals (a 0.05% relative increase), their study has a very high statistical power and would not run into substantial type M error. A retrospective analysis is thus a very convenient way to think about the statistical power of a study to accurately detect alternative effect sizes.

3.2 Standard Epidemiology Literature

Hundreds of papers have been published on the short-term health effects of air pollution in epidemiology, medicine and public health journals. A large fraction of articles are based on Poisson generalized additive models, which allow to flexibly adjust for the temporal trend of health outcomes and for non-linear effects of weather parameters. This literature spans over 20 years and has replicated analyses in a large number of settings, providing crucial insights on the acute health effect of air pollution. Advocates of causal methods would surely argue that these articles could suffer from omitted variable biases. Even if they may be more biased, we find it valuable to assess their potential statistical power issues and compare them with causal inference papers.

To gather a corpus of relevant articles, we use the following search query on PubMed and Scopus to select studies on the short-term health effects of air pollution:

```
'TITLE(("air pollution" OR "air quality" OR "particulate matter" OR "ozone"',
'OR "nitrogen dioxide" OR "sulfur dioxide" OR "PM10" OR "PM2.5" OR', ' "carbon
dioxide" OR "carbon monoxide")', 'AND ("emergency" OR "mortality" OR "stroke"
```

```
OR "cerebrovascular" OR', '"cardiovascular" OR "death" OR "hospitalization")',
'AND NOT ("long term" OR "long-term")) AND "short term"'
```

We retrieve the abstracts of 1834 articles. We then extract estimates and confidence intervals from these abstracts using REGular EXpressions. We illustrate this procedure using one sentence of a randomly selected article from this literature review (Vichit-Vadakan et al. 2008):

> "The excess risk for non-accidental mortality was **1.3% [95% confidence interval (CI), 0.8–1.7]** per 10 $\mu g/m^3$ of PM10, with higher excess risks for cardiovascular and above age 65 mortality of **1.9% (95% CI, 0.8–3.0)** and **1.5% (95% CI, 0.9–2.1)**, respectively."

Our algorithm detects phrases such as "95% confidence interval (CI)" or "95% CI" and looks for numbers directly before this phrase or after and in a confidence interval-like format. Using this method, we retrieve 2666 estimates from 784 abstracts. We then read these abstracts and filter out articles whose topic falls outside of the scope of our literature review. Our corpus is thus composed of 668 articles for which we detect 2155 estimates. Importantly, the set of articles considered is limited to those displaying confidence intervals and point estimates in their abstracts.

Based on this subset of articles, we implement a retrospective analysis in which we check the overall sensitivity of studies for true effect sizes expressed as fraction of observed estimates. Without carefully reading each article, we cannot make more informed guesses about true effect sizes since estimates are expressed for different increases in air pollution concentration. We think that our rough approach is still valuable since a well-designed study should be able to detect effect sizes smaller than the estimated one. For instance, if we find that a study has a statistical power of 30% when we assume that the true effect is 3/4 of the measured estimate, it is likely that the study is not very robust to statistical power issues.

Our results for the standard literature are at first sight reassuring. If the true effect sizes of the studies were equal to 75% of estimated coefficients, the median statistical power would be equal to 85% and the median exaggeration factor would be only 1.1. At least 50% of this literature does not seem to suffer from substantial statistical power issues. Type S error is not an issue for most articles. Yet, even if the measured effect was close to the true effect, a non-negligible proportion of articles would display low statistical power and presents a substantial risk of making a type M error. About 47% of estimates would not reach the conventional 80% statistical power threshold if the true effect was 75% the size of the measured effect. Concernedly, for these under-powered studies, the average type M is 1.9 and the median 1.5. We also observe that the proportion of under-powered studies has been stagnating since the 1990s, revealing that practices regarding statistical power have not evolved over time.

Finally, skeptic researchers could rightly complain that assuming for each study a true effect size equal to 75% of the estimate is arbitrary. To overcome this criticism, we expand our review of the standard epidemiology literature by running statistical power calculations based on two recent meta-analyses: one by Shah et al. (2015) on mortality and emergency admission for stroke, and the other one by Orellano et al. (2020) on broader causes of mortality. We use the meta-analysis estimates as true effect sizes for the 290 studies gathered by Shah et al. (2015) and (Orellano et al. 2020). This is the approach recommended by Gelman and Carlin (2014) and Ioannidis et al. (2017) to make more informed guesses about true effect sizes. 60% of studies in Orellano et al. (2020) have a statistical power below 80%. The median exaggeration ratio of statistically significant estimate is equal to 2. The proportion of under-powered studies is similar in Shah et al. (2015) but the median type M error is equal to 3. With more informed guesses about true effect sizes, we clearly see under-powered studies are an issue in the standard epidemiology literature.

3.3 Causal Inference Literature

Using Google Scholar, PubMed, we search papers using causal inference methods and investigating the short-term effects of air pollution on mortality or emergency admission outcomes. Specifically, we only consider articles that exploit short-run exogenous shocks such as air pollution alerts, public transport strikes, changes in wind direction, thermal inversions, to name but a few. For instance, we did not select articles on the impact of low emission or congestion pricing zones as they evaluate health effects over several months or years. In Table 2, we display the 29 articles that match our search criteria. We read each article and retrieve the estimates and standard errors for the main results: for simplicity, we only select one of the main results discussed by the researchers. We also record the numbers of observations and summary statistics on the outcome and independent variables to compare studies by standardizing the estimated effect sizes.

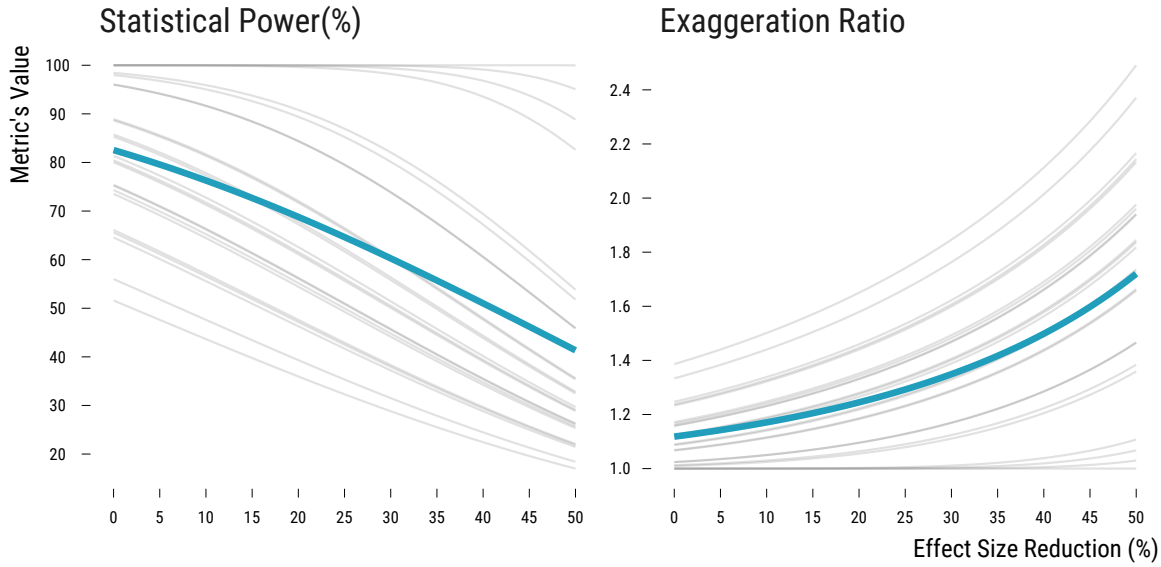
Table 2: Our Corpus of Papers from the Causal Inference Literature.

Article	Location	Health Outcome	Independent Variables	Study Design
Arceo et al. (2016)	Mexico City, Mexico	Infant Mortality	PM10, Thermal Inversion (IV)	Instrumental Variable
Austin et al. (2020)	Counties, USA	Rates of Confirmed COVID-19 Deaths	PM2.5 (air pollutant), Wind Direction (IV)	Instrumental Variable
Baccini et al. (2017)	Milan, Italy	Non-Accidental Mortality	Dummy for PM10 Concentration >To 40 $\mu g/m^3$	Propensity Score Matching
Barwick et al. (2018)	All Cities, China	Number of Health Spending Transactions	PM2.5, Spatial Spillovers of PM2.5 (IV)	Instrumental Variable
Bauernschuster et al. (2017)	5 Largest Cities, Germany	Admissions for Abnormalities of Breathing (age below 5)	PM10, Public Transport Strikes Dummy	Difference in Differences
Beard et al. (2012)	Salt Lake County, USA	Emergency Visits For Asthma	Thermal Inversions	Time-stratified case-crossover design
?	Toronto, Canada	Asthma-Related Emergency Department Visits	Air Quality Eligibility, Air Quality Altert	Fuzzy Regression Discontinuity
Deryugina et al. (2019)	Counties, USA	All Causes of Mortality (Age 65+)	PM2.5, Wind Direction (IV)	Instrumental Variable
Ebenstein et al. (2015)	2 Cities, Israel	Hospital Admissions Due To Lung Illnesses	PM10 (air pollutant), Sandstorms (IV)	Instrumental Variable
Forastiere et al. (2020)	Milan, Italy	Non-Accidental Mortality	Setting PM10 Daily Exposure Levels >To 40 $\mu g/m^3$ To 40	Generalized Propensity Score
Giaccherini et al. (2021)	Municipalities, Italy	Respiratory Hospital Admission	PM10, Public Transport Strikes	Difference in Differences
Godzinski and Suarez Castillo (2019)	10 Cities, France	Emergency Admissions for Upper Respiratory System (Age 0-4)	CO, Public Transport Strikes	Difference in Differences
Halliday et al. (2019)	Hawaii, USA	ER Admission for Pulmonary Outcomes	PM2.5, SO2 Emissions From Kilauea Volcano and Wind Direction (IV)	Instrumental Variable
He et al. (2016)	34 Urban Districts, China	Monthly Standardized Mortality Rate	PM10, Regulation and Traffic Control Status (IV)	Instrumental Variable
He et al. (2020)	China	Monthly Number of Deaths for All-Causes	PM2.5, Straw Burning (IV)	Instrumental Variable
Isphording and Pestel (2021)	Counties, Germany	Mortality of Covid-19 Positive Male Patients (Age 80+)	PM10, Wind direction (IV)	Instrumental Variable
Jans et al. (2018)	Sweden	Children Health Care Visits for Respiratory Illness	PM10, Thermal Inversion (IV)	Instrumental Variable
Jia and Ku (2019)	South Korea	Mortality Rates for Respiratory and Cardiovascular Diseases	Dusty Days Times China's AQI	Reduced-Form
Kim et al. (2013)	South Korea	Hospital Admissions for Respiratory Illnesses	PM10 (air pollutant), Average PM10 Level By Date (IV)	Instrumental Variable
Knittel et al. (2016)	California, USA	Infant Mortality	PM10, Road Traffic Flow and Weather variables (IV)	Instrumental Variable
Moretti and Neidell (2011)	South California, USA	Hospital Admissions for Respiratory Illnesses	O3, Vessel Traffic (IV)	Instrumental Variable
Mullins and Bharadwaj (2015)	Santiago Metropole, Chile	Cumulative Deaths (age >64)	PM10, Air quality Alerts	Matching + Difference in Differences
Schlenker and Walker (2016)	California, USA	Acute Respiratory Hospitalization	CO, Planes Taxi Time (IV)	Instrumental Variable
Schwartz et al. (2015)a	Boston, USA	Non-Accidental Mortality	PM2.5, Back Trajectories of PM2.5 (IV)	Instrumental Variable
Schwartz et al. (2017)	Boston, USA	Non-Accidental Mortality	PM2.5, Height Of Planetary Boundary Layer and Wind Speed (IV)	Instrumental Variable
Schwartz et al. (2018)	135 Cities, USA	Non-Accidental Mortality	PM2.5, Planetary Boundary Layer, Wind Speed, and Air Pressure (IV)	Instrumental Variable
Sheldon and Sankaran (2017)	Singapore	Acute Upper Respiratory Tract Infections	Pollutant Index, Indonesian Fire Radiative Power (IV)	Instrumental Variable
Williams et al. (2019)	USA	Asthma Rescue Event	PM2.5	Poisson fixed-effects models
Zhong et al. (2017)	Beijing, China	Ambulance Call Rate for Coronary Heart Problem	NO2, Number 4 Day (IV)	Instrumental Variable

Notes: For each study, we report its location, one of the health outcome analyzed, the independent variables (the air pollutant and in the case of an instrumental variable strategy, the instrument) and the study design.

To evaluate potential statistical power issues in this literature, we first proceed exactly as for the standard epidemiology literature. We compute the statistical power, the exaggeration factor and the probability to get an estimate of the wrong sign for all studies based on hypothetical true effect sizes expressed as decreasing fraction of observed estimates. In Figure 4, each gray line represent the statistical power and average type M error curves of an article. The blue lines represent the average power and exaggeration factor of all causal inference studies.

Figure 4: Statistical Power and Type M Error of Causal Inference Studies.



Notes: For each causal inference paper, we compute its statistical power and the average type M error for decreasing effect sizes expressed as percentage reduction in observed estimates. Each gray line represents a specific causal inference paper. The blue lines are the average of a metric for all causal inference papers.

If the true effect size of each study was equal to 75% of the estimate, the median statistical power would be about to 60% and the median Type M error would be 1.3. In the causal inference literature, at least half of studies have enough statistical power so that statistical significant estimates are not inflated. In Figure 4, we can however see that there is a wide heterogeneity in the robustness of studies to statistical power issues—some of them are relatively well powered while others run quickly into Type M error. A large share of studies in the literature would not

have designs with enough statistical power to detect effects of half the size of their observed estimates. In that scenario, the median statistical power would be about 40% and the median type M error would be 1.8. Overall, this comprehensive retrospective analysis of the literature reveals that some studies are under-powered and could run into type M error. It may help explain why there is a large heterogeneity in effect sizes across articles.

Again, expressing true effect sizes as decreasing fraction of observed estimates is arbitrary. To overcome this limit, we carry out another retrospective analysis where we take as true effect sizes the estimates that would be predicted using non-causal inference methods. We do so for the subset of the 9 instrumental variables that also display estimates in the case when the air pollutant concentration is not instrumented. Two reasons are often advanced in the causal literature to explain the discrepancy between instrumented and non-instrumented estimates: (i) instrumental variables help overcoming omitted variable bias and (ii) if the air pollution is measured with classical error, instrumental variables also reduce the resulting attenuation bias. We think that, for some studies, statistical power issues could also partly explain the observed difference between causal and non-causal methods. In Table 2, we display the statistical power, the average type M error and the probability to make a type S error for instrumental variable studies. For some studies, the statistical power of the instrumental variable strategy could be extremely low. This results in large type M errors, which magnitude partially close the gap between instrumented and non-instrumented estimates. Given this possibility, future research should carry out quantitative bias analysis to explore the trade-off between using an instrumental variable strategy to overcome omitted variable and attenuation biases and running into a type M error due to low statistical power (Rosenbaum 2010, Dorie et al. 2016, VanderWeele and Ding 2017, Cinelli and Hazlett 2020).

Table 3: A Retrospective Analysis of Instrumental Variable Papers.

Paper	Power (%)	Type S Error (%)	Type M Error
Giaccherini et al. (2021)	5	43.3	40.7
Halliday et al. (2019)	6	16.6	6.9
Schlenker and Walker (2016)	7	13.8	6.1
Moretti and Neidell (2011)	11	3.7	3.5
Arceo et al. (2016)	12	2.4	3.1
Barwick et al. (2018)	23	0.3	2.1
Deryugina et al. (2019)	34	0.1	1.7
Ebenstein et al. (2015)	52	0	1.4
Schwartz et al. (2018)	64	0	1.3

Notes: For each study based on an instrumental variable strategy, we computed the statistical power, the average type M error and the probability to make a type S error using the non-instrumented estimate as a guess for the true effect size.

4 Prospective Analysis of Causal Inference Methods

The review of the standard epidemiology and causal literatures shows that some articles could have produced inflated estimates on the short-term health effects of air pollution. This analysis however does not allow us to clearly identify which parameters of a study influence its statistical power. We therefore implement a prospective analysis to overcome this limitation (Gelman and Carlin 2014, Altoè et al. 2020). We run simulations based on real-data to emulate the main empirical strategies found in the literature. Using real-data avoid us the difficult task to model the long-term and seasonal variations in health outcomes but also the specific effects of weather variables such as temperature. We first explore how statistical power is related to the treatment effect size, the number of observations, the proportion of treated units and the distribution of the health outcome. We then try to replicate the design of flagship publications to highlight their potential weaknesses with respect to low statistical power issues.

In this section, we describe how we implement these simulations. We start by presenting the research designs we emulate, then briefly describe the data we rely on and finally detail our simulation procedure.

4.1 Research Designs Emulated

Several empirical strategies have been implemented to estimate the short-term health effects of air pollution. In our simulations, we try to emulate the main ones found in the literature. The standard strategy consists in directly estimating the dose-response between an air pollutant and an health outcome. In the epidemiology literature, researchers often rely on Poisson generalize additive models where the daily count of an health outcome is regressed on the concentration of an air pollutant, while flexibly adjusting for weather parameters, seasonal and long-term variations. Because most causal methods are estimated with linear regression, our simulations are instead based on ordinary least square estimation to approximate the warhorse model used by epidemiologists.

The standard strategy could however be prone to omitted variable bias and measurement error. A growing number of articles therefore exploit exogenous variations in air pollution. Most causal inference papers rely on instrumental variable designs where the concentration of an air pollutant is instrumented by thermal inversions (Arceo et al. 2016, Jans et al. 2018), wind patterns (Schwartz et al. 2018, Deryugina et al. 2019, Isphording and Pestel 2021), extreme natural events such as sandstorms or volcano eruptions (Ebenstein et al. 2015, Halliday et al. 2019), or variations in transport traffic (Moretti and Neidell 2011, Knittel et al. 2016, Schlenker and Walker 2016). In our simulations, we simplify the instrument variable strategy by considering only binary instruments, such as the presence of a thermal inversion or not. Besides, the occurrence of exogenous shocks are completely random.

We also emulate two other empirical strategies found in the causal inference literature. The first one consists in reduced-form or difference-in-differences strategies where researchers do not instrument the impact of exogenous shocks on air pollution. These articles mostly focus on public transport strikes (Bauernschuster et al. 2017, Godzinski and Suarez Castillo 2019, Giaccherini et al. 2021). Again, we make the simplifying assumption that these events are completely random in our simulations. We do not model the resulting air pollution increases but only focus on the impact of these shocks on health outcomes. The second strategy concerns the analysis of air quality alerts using regression-discontinuity design (Chen et al. 2018). For simplification, we only model sharp designs where an air quality is always activated above a randomly chosen threshold.

4.2 Data

Our simulation exercises are based on a subset of the US National Morbidity, Mortality, and Air Pollution Study (NMMAPS). The dataset has been exploited in several major studies of the early 2000s to measure the short-term effects of ambient air pollutants on mortality outcomes (Peng and Dominici 2008). It is openly available and allows us to work with increasing sample sizes for our simulations. Specifically, we extracted daily data on 68 cities over the 1987-1997 period, which represent 4,018 observations per city, for a total sample size of 273,224 observations. For each city, the average temperature (C°), the standardized concentration of carbon monoxide (CO), and mortality counts for several causes are recorded. We choose to work with CO as it is the air pollutant measured in most cities over the period. Less than 5% of carbon monoxide concentrations and average temperature readings are missing in the initial data set and we impute them using the chained random forest algorithm provided by the `missRanger` package (Mayer 2019).

4.3 Simulations Set-Up

Our simulation procedure follows 7 main steps:

1. Draw a study period and a sample of cities.
2. For instrumental variable, reduced-form and regression-discontinuity designs, randomly allocate days to exogenous shocks.
3. Create the counterfactual health outcome based on the treatment effect size.
4. Run the model of the empirical strategy.
5. Store the point estimate of interest and its standard error.
6. Repeat the procedure 1000 times.
7. Finally compute the statistical power, the exaggeration ratio of statistically significant estimates and the probability that they are of the wrong sign.

In the first step, a study period is drawn at random. Then, a given number of cities and days are sampled from the data. We consider the same study period for each city. The second step only concerns causal inference methods. The drawing procedure for days exposed to exogenous shocks is specific to the inference strategy and the proportion of treated observations desired. For instrumental variable and reduced-form strategies, the treatment status of each day is drawn from a Bernoulli distribution with parameter equal to the proportion of exogenous shocks. For air pollution alerts, we randomly draw a threshold from a uniform distribution and select a bandwidth such that it yields the correct proportion of treated observations. In the third step, we add to the data the treatment effect size of air pollution or the direct effect of an exogenous shock on an health outcome. For the reduced-form and regression discontinuity designs, we follow the Neyman-Rubin causal framework by creating a Science table (Rubin 1974). The observed values of an health outcome in the dataset represent the potential outcomes of days when they are not exposed to the treatment. To create the counterfactuals, we add a treatment effect

drawn from a Poisson distribution with parameter corresponding to the effect size. For the standard regression approach, we first estimate the model and then generate fake observations of an health outcome based the estimated coefficients (Peng et al. 2006). The treatment effect size is added to the data by modifying the value of the air pollution coefficient. For the instrumental variable strategy, we use the same method as for the standard regression approach but add in the first-stage the effect size of the instrument on the air pollutant. We then estimate a two-stage least squares model, modify the coefficient for the effects of the air pollutant on an health outcome, and finally generate the fake observations of the health outcome.

5 Results

In this section, we first display how statistical power and the inflation of statistically significant estimates evolve with each study's parameter. We then provide results on statistical power issues for several flagship publications.

5.1 Evolution of Power, Type M and S Errors with Study Parameters

First, we analyze how statistical power, type M and S errors are affected by the value of different study parameters. To do so, we set baseline values for these parameters and vary the value of each of them one by one. This enables us to get a sense of the impact of each parameter, other things being equal. The baseline parameters are such that:

- The sample size is equal to 100,000 observations (2500 days × 40 cities).
- The effect size of air pollution or an exogenous shock is equal to a 1% relative increase in an health outcome.

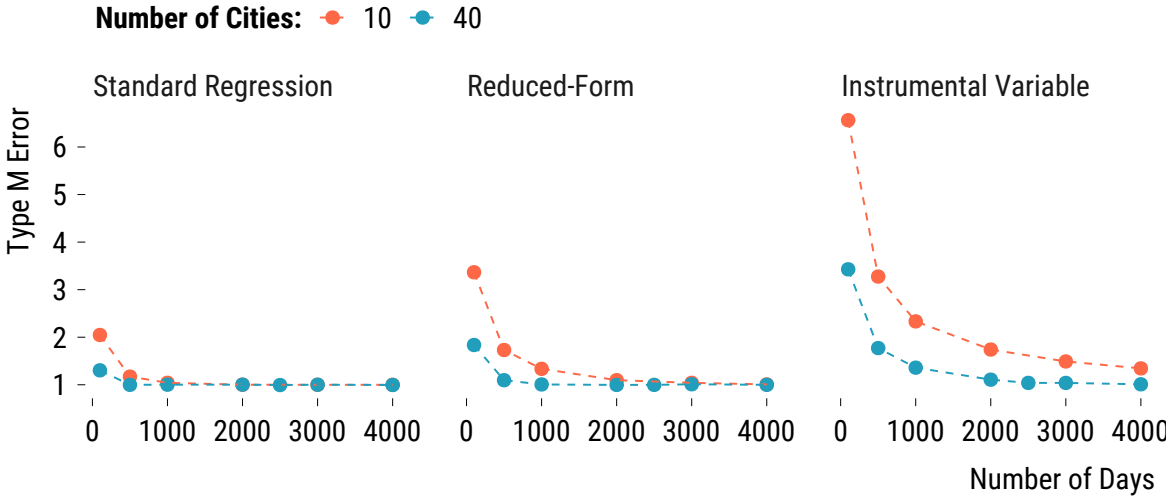
- The proportion of exogenous shocks represents 50% of observations. For air pollution alerts analyzed with regression discontinue designs, we choose a smaller proportion of treated units: 10%.
- The health outcome is the total daily number of non-accidental deaths. It is the health outcome with the largest average number of counts—the average daily mean is equal to 23 cases.

For all statistical models, we adjust for temperature, temperature squared, city and calendar (weekday, month, year, month×year) fixed effects. We also repeat the simulations for a smaller sample size of 10,000 observations.

Sample Size

As shown in Figure 5, we obviously find that, for all identification methods, statistical power increases and type M error decreases with the number of observations.

Figure 5: Evolution of Type M Error against Sample Size.



Notes: The true effect size is a 1% relative increase in the health outcome. The health outcome used in the simulations is the total number of non-accidental deaths. The proportion of exogenous units is 50% for instrumental variable and reduced-form designs.

Yet, statistical power and type M error issues arise even for a large number of observations. For a sample size of 40,000 observations, an instrumental variable strategy

would only have a statistical power of 54% and would overestimate the true effect by a factor of 1.4. On the contrary, a standard regression strategy is much less prone to power issues than the instrumental variable strategy. This is explained by the fact that the variance of the two stage least-square estimator is larger than the variance of the ordinary least square estimator. In our simulations, we also note that, for all identification method, Type S error is not a problem for any sample sizes.

Effect Size

The second unsurprising result of our simulations is that the larger the effect size, the larger the power and the lower type M and S errors are. With our advantageous baseline parameters, statistical power issues however start to appear in instrumental variable and regression discontinuity designs for effect sizes below 1%. For instance, for an effect of 0.5%, the average type M error is about 1.7. Such effect sizes are similar to those sometimes found in the standard epidemiology literature. As for results on sample sizes, standard regression and reduced-form strategies suffer less from power issues, even for small effects.

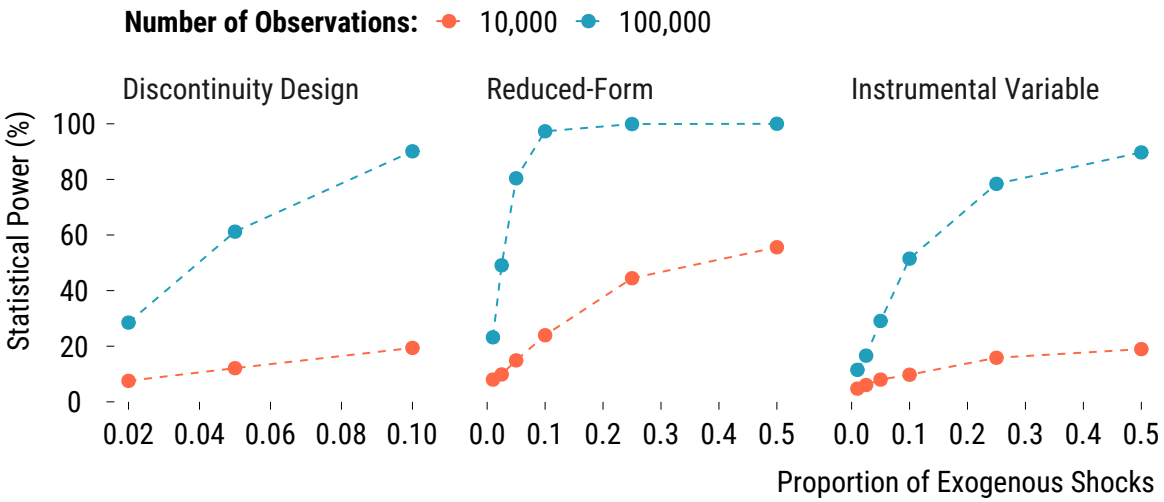
Proportion of Exogenous Shocks

The link between the proportion of exogenous shocks and statistical power might be less known to researchers. In Figure 6, we see that the statistical power increases with higher proportions of treated units for instrumental variable, regression discontinuity and reduced-form designs. As in the case of randomized controlled trials, the precision of studies will we be maximized when half of the observations are exposed to the treatment of interest.

Conversely, as shown in Figure 7, the average Type M error increases as the proportion of exogenous shocks decreases.

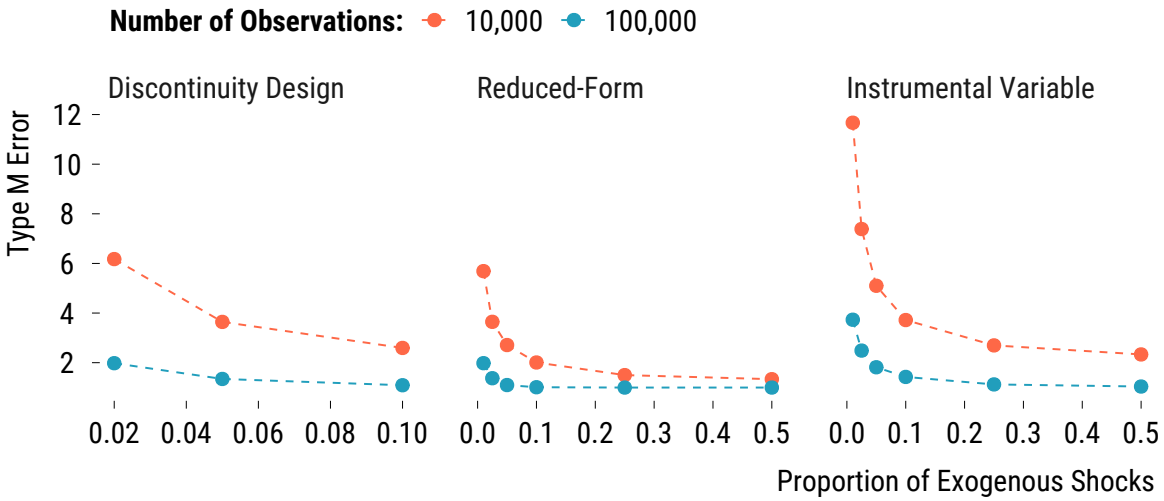
Air pollution alerts, thermal inversion or transportation strikes are however rare

Figure 6: Evolution of Statistical Power with the Proportion of Exogenous Shocks.



Notes: The true effect size is a 1% relative increase in the health outcome. The health outcome used in the simulations is the total number of non-accidental deaths.

Figure 7: Evolution of Type M Error with the Proportion of Exogenous Shocks.



Notes: The true effect size is a 1% relative increase in the health outcome. The health outcome used in the simulations is the total number of non-accidental deaths.

events. They can represent less than 5% of the observations in some studies. With a dataset of 10,000 observations, the average type M error is 2.7 for reduced-form strategies. The causal inference literature might therefore be particularly prone to type M error due to a very low proportion of treated units, even though sample sizes

are often large.

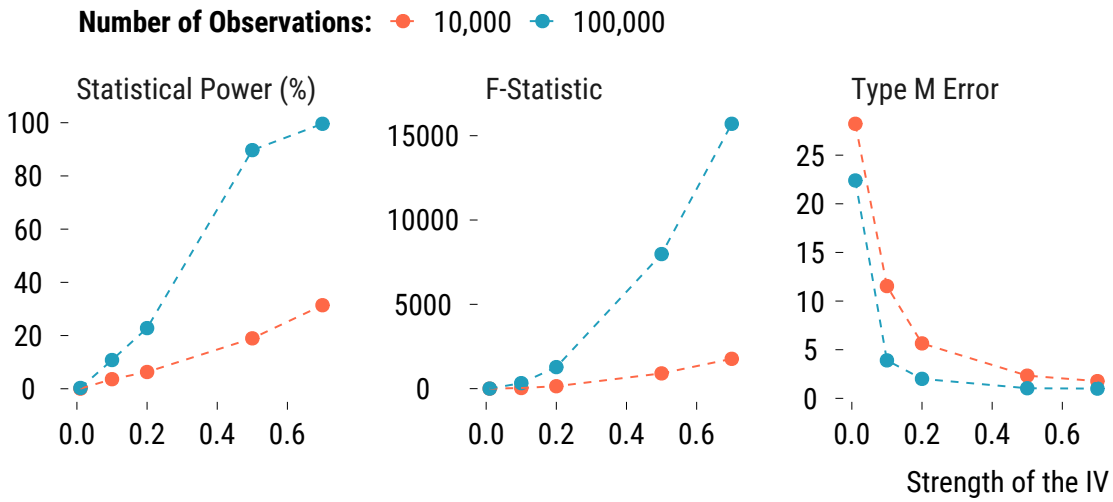
Average Count of Cases of the Health Outcome

Perhaps less known to economists than the influence of sample and effect sizes, the average count of cases also critically affects statistical power. For instance, a 1% increase in the number of deaths in a setting where there are only 2 deaths per day corresponds to rare additional deaths that might therefore be more difficult to detect. To emulate situations with various number of cases, we consider three different outcome variables, with different counts of cases: the total number of non-accidental deaths (daily mean $\simeq$ 23), the total number of respiratory deaths (daily mean $\simeq$ 2) and the number of chronic obstructive pulmonary disease cases for people aged between 65 and 75 (daily mean $\simeq$ 0.3). With baseline parameters and in the case of the large dataset, we find that statistical power is close to 100% when empirical strategies target a 1% increase in the total number of non-accidental deaths. However, statistical power quickly drops when the average count of cases decreases. For instance, an instrumental variable strategy has only 16% of statistical power to detect an increase by 1% in respiratory deaths. The average type M error is then equal to 2.4. For chronic obstructive pulmonary deaths, the situation is even worse, with an average type M error of 5.9. Studies with a small count of cases may therefore lead to extreme statistical power issues.

Issues Specific to the Instrumental Variable Design

For instrumental variable strategies, we also analyze how the statistical power is affected by the strength of the instrument. In our simulations, we define the strength of the instrument as the standardize effect size on the air pollutant concentration. A strength equals to 0.2 means that the instrument increases the concentration by 0.2 standard deviation.

Figure 8: Evolution of Type M Error with the Proportion of Exogenous Shocks.



Notes: The true effect size is a 1% relative increase in the health outcome. The health outcome used in the simulations is the total number of non-accidental deaths. Half of the observations are exposed to exogenous shocks. The instrument increases the air pollution concentration by 0.5 standard deviation.

As visible in Figure 8, we find that statistical power collapses and type M error soars when the instrument's strength decreases. Importantly, this issue arises for rather large instrument's strengths. Even in the case of the large data set with 100,000 observations, an instrumental variable's strength of 0.2, and effect size of a 1% increase in the health outcome, statistical power is only 23% and the average type M error is 2. This statistical power issue arises for a F-statistics of 1278! A large F-statistic could therefore hide a weak instrumental variable that results in a low statistical power.

5.2 Simulating Flagship Studies

The simulation results of the previous section help build the intuition for the parameters influencing the statistical power of studies. Yet, they represent an ideal setting, with relatively large sample size, proportion of treated units, outcome counts and instrumental variable strength. These parameters may not perfectly represent

actual studies. For each causal inference method, we therefore consider a realistic set of parameters based on examples from the literature. We then vary the value of key parameters one by one in order to see what could be changed in each study to avoid running into power issues.

Public Transport strikes

Public transport strikes are unique but rare exogenous events where air pollution increases. Even in a large data set, with several cities and a long study period, the proportion of treated days might be very small. For instance, Bauernschuster et al. (2017) investigate the effect of public transport strikes on air pollution and emergency admission in the five biggest German cities over a period of 6 years. The sample size of the study is equal to 11,000 observations but there are only 45 1-day strikes. This study could be prone to statistical power issues since the proportion of treated units is 0.4%. We thus try to simulate with our data a similar design. In our baseline simulation, we set as the true effect size the point estimate found by Bauernschuster et al. (2017): days with strikes see an 11% relative increase in the health outcome of interest. The average count of cases for our health outcome—the total number of respiratory deaths–is however 3 times larger than the one in their study, which is equal to 0.69.

In the baseline scenario, we find that the statistical power is only 15% and the average type M error is 2.7. If the researchers have looked at the effect for an health outcome with an average of 23 cases per day, there would however be no statistical power issues. The effect size found by the authors could nonetheless be argued to be a very large increase in an health outcome. If the true effect was only 5% and the average count of the health outcome was 23, there would still be a substantial risk to overestimate statistically significant estimates by a factor of 1.8! Estimating accurately the effects of rare exogenous events on health outcomes with few cases

seem therefore difficult.

Air pollution Alerts

Air pollution alerts are also rare events. Contrary to public transport strikes or thermal inversions, their effects are estimated using regression discontinuity design. Only observations closed to the air quality threshold are included in the analysis. As a consequence, the effective sample size may end up being particularly small. For instance, in Chen et al. (2018), while the initial sample size is equal to 3652 observations, the effective sample size is only of 143 (100 control observations and 43 treated ones). The proportion of treated observation is 1.2%. With our data, we try to approximate the setting of Chen et al. (2018). In the baseline scenario, we sample one city with a time period of 3652 days and randomly allocate to treatment 1.2% of observations. We also consider a true effect size of 12%, as found in the study. The average number of cases of their health outcome is 26 cases per day. In our simulations, we use the total number of non-accidental deaths as our outcome variable since the daily mean is equal to 23 deaths.

In the baseline scenario, we find that the statistical power is only 10% and the average type M error is 4.6. If we consider smaller true effect sizes, type M error shoots up and power collapses. As a consequence, we cannot put too much confidence in reported estimates from studies with such a small sample size and few air quality alerts.

Instrumenting Air Pollution

Finally, we investigate the most common strategies used in the causal inference literature, which are based on instrumental variables. These papers often present very large data sets. For instance Schwartz et al. (2018) gathered 591,570 observations (135 cities with a length of study of approximately 4382 days). In this study, air

pollution is instrumented with a complex mix of variables and we cannot easily observe the proportion of treated units. The effect size found by the authors is equal to a 1.5% relative increase in an health outcome with an average daily number of cases equal to 23. In our simulations, we therefore assume that half of the observations are exposed to exogenous shocks. We only vary the strength of the instrument and use the total number of non-accidental deaths as the outcome variable. Our data set being smaller than the one used in the study, we only consider 2500 days and 40 cities.

If the instrumental variable increases air pollution concentration by 0.5 standard deviation, we find a statistical power of nearly 100% and an average type M error of 1. Yet, for smaller values of the instrument's strength, statistical power rapidly decreases. For an instrument's strength of 0.2, the statistical power is 48% and the average type M error is 1.4. For a strength of 0.1, power is only 16% and the average type M error is 2.6! In these two scenarios, the values of the F-statistic remain extremely large, with respective values equal to 1287 and 320. A large F-statistic can be a poor indicator of statistical power issues.

6 Discussion

> *"I think that when we know that we actually do live in uncertainty, then we ought to admit it."*
>
> — Richard P. Feynman

Our findings should make us worried about statistical power issues when we are trying to estimate the acute health effects of air pollution. Our retrospective analysis of the literature suggests that under-powered studies with inflated effect sizes could be an actual issue both in the standard epidemiology and the causal

inference literatures. We thus recommend to adopt retrospective calculations since they are very easy to implement and force us to reflect on the range of plausible effect sizes we are trying to estimate.

Unfortunately, a retrospective analysis will not help researchers understand which parameters of the research design influence the statistical power of their studies. Our prospective analysis, using simulations based on real-data, fills this gap and leads to issue 4 warnings. First, sample size matters for all causal inference methods but especially for the regression-discontinuity design applied to air pollution alerts. Given the sample size its entails, we advice researchers to interpret findings with extra-care as type M error can be extremely large, even for a large guess about the true effect size. Second, despite their large sample sizes, researchers exploiting rare exogenous shocks such as transport strikes should be aware that the small proportion of exogenous shocks observed in their studies can lead to a dramatically low statistical power. Third, researchers are likely aware that two-stage least square estimates are inherently less precise than ordinary least square estimates. It however makes instrumental variable strategies more prone to type M error. If one thinks that omitted variable and attenuation biases are small, the benefits of using an instrumental variable strategy could be questioned. The trade-off between targeting an unbiased estimate with causal inference method and the risk of running into a type M error should be a fruitful area of research for quantitative bias analysis (Rosenbaum 2010, Dorie et al. 2016, VanderWeele and Ding 2017, Cinelli and Hazlett 2020). Fourth, the power of all research designs in the literature is driven by the average count of the health outcome. Many articles investigate the acute effects of air pollution for specific groups such as children and the elderly. In such settings, there is potentially a huge risk to make a type M error, even with large sample sizes. While they are more involved than a retrospective analysis, simulating the research design researchers want to implement is the best way to assess if it

could suffer from statistical power issues. Our simulation codes in the replication material provide researchers a template to run such prospective analysis.

On top of these specific warnings, we insist that the literature would benefit from reforming editors and researchers' attitude towards statistically insignificant results Ziliak and McCloskey (2008a), Wasserstein and Lazar (2016), McShane et al. (2019). The null hypothesis testing framework remains very strong in the field, especially for causal inference papers since nearly all of them dichotomize evidence using the 5% significance threshold (Greenland 2017). This statistical significance filter leads to publication bias and is at the very heart of the inflation statistically significant estimates in under-powered studies (Amrhein et al. 2019, Gelman et al. 2020, Romer 2020). Even if researchers could not improve the statistical power of their studies, the distribution of the acute health effects of air pollution could be therefore be more accurate if statistically insignificant results are not kept in the file drawer.

We finally hope that our article reminds us that a credible identification strategy does not necessarily lead to a correct estimation of the actual true effect (Young 2019). Published results are not carved in marble: when researchers qualify estimates as "statistically significant", there is often much more uncertainty lying behind, an uncertainty that should be computed and embraced to better help policymakers evaluate the adverse effects of air pollution. Retrospective and prospective analyses can push the literature forward.

References

Allaire, JJ and Iannone, Rich and Presmanes Hill, Alison and Xie, Yihui. Distill for r markdown, 2018. URL https://rstudio.github.io/distill.

Altoè, Gianmarco and Bertoldo, Giulia and Zandonella Callegher, Claudio and Tof-

falini, Enrico and Calcagnì, Antonio and Finos, Livio and Pastore, Massimiliano. Enhancing Statistical Inference in Psychological Research via Prospective and Retrospective Design Analysis. *Frontiers in Psychology*, 10:2893, January 2020. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.02893.

Amrhein, Valentin and Greenland, Sander and McShane, Blakeley B. Statistical significance gives bias a free pass. *European journal of clinical investigation*, 49 (12):e13176, 2019.

Arceo, Eva and Hanna, Rema and Oliva, Paulina. Does the Effect of Pollution on Infant Mortality Differ Between Developing and Developed Countries? Evidence from Mexico City. *The Economic Journal*, 126(591):257–280, March 2016. ISSN 00130133. doi: 10.1111/ecoj.12273.

Austin, Wes and Carattini, Stefano and Mahecha, John Gomez and Pesko, Michael. Covid-19 mortality and contemporaneous air pollution. Technical report, CESifo Working Paper, 2020.

Baccini, Michela and Mattei, Alessandra and Mealli, Fabrizia and Bertazzi, Pier Alberto and Carugno, Michele. Assessing the short term impact of air pollution on mortality: A matching approach. *Environmental Health*, 16(1):7, December 2017. ISSN 1476-069X. doi: 10.1186/s12940-017-0215-7.

Barwick, Panle Jia and Li, Shanjun and Rao, Deyu and Zahur, Nahim Bin. The Morbidity Cost of Air Pollution: Evidence from Consumer Spending in China. Technical Report w24688, National Bureau of Economic Research, Cambridge, MA, June 2018.

Bauernschuster, Stefan and Hener, Timo and Rainer, Helmut. When Labor Disputes Bring Cities to a Standstill: The Impact of Public Transit Strikes on Traffic, Accidents, Air Pollution, and Health. *American Economic Journal: Economic Policy*, 9(1):1–37, February 2017. ISSN 1945-7731, 1945-774X. doi: 10.1257/pol.20150414.

Beard, John D. and Beck, Celeste and Graham, Randall and Packham, Steven C. and Traphagan, Monica and Giles, Rebecca T. and Morgan, John G. Winter Temperature Inversions and Emergency Department Visits for Asthma in Salt Lake County, Utah, 2003–2008. *Environmental Health Perspectives*, 120(10):1385–1390, October 2012. ISSN 0091-6765, 1552-9924. doi: 10.1289/ehp.1104349.

Bell, Michelle L and Samet, Jonathan M and Dominici, Francesca. Time-series studies of particulate matter. *Annu. Rev. Public Health*, 25:247–280, 2004.

Bhaskaran, Krishnan and Gasparrini, Antonio and Hajat, Shakoor and Smeeth, Liam and Armstrong, Ben. Time series regression studies in environmental epidemiology. *International journal of epidemiology*, 42(4):1187–1195, 2013.

Bind, Marie-Abèle. Causal Modeling in Environmental Health. *Annual Review of Public Health*, 40(1):23–43, April 2019. ISSN 0163-7525, 1545-2093. doi: 10.1146/annurev-publhealth-040218-044048.

Black, Bernard and Hollingsworth, Alex and Nunes, Leticia and Simon, Kosali. Simulated power analyses for observational studies: An application to the affordable care act medicaid expansion. Technical report, National Bureau of Economic Research, 2019.

Brodeur, Abel and Lé, Mathias and Sangnier, Marc and Zylberberg, Yanos. Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, 8 (1):1–32, 2016.

Brodeur, Abel and Cook, Nikolai and Heyes, Anthony. Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics. *American Economic Review*, 110(11):3634–3660, November 2020. ISSN 0002-8282. doi: 10.1257/aer.20190687.

Button, Katherine S and Ioannidis, John PA and Mokrysz, Claire and Nosek, Brian A and Flint, Jonathan and Robinson, Emma SJ and Munafò, Marcus R. Power failure: why small sample size undermines the reliability of neuroscience. *Nature*

reviews neuroscience, 14(5):365–376, 2013.

Camerer, Colin F and Dreber, Anna and Holzmeister, Felix and Ho, Teck-Hua and Huber, Jürgen and Johannesson, Magnus and Kirchler, Michael and Nave, Gideon and Nosek, Brian A and Pfeiffer, Thomas and others. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644, 2018.

Chen, Hong and Li, Qiongsi and Kaufman, Jay S and Wang, Jun and Copes, Ray and Su, Yushan and Benmarhnia, Tarik. Effect of air quality alerts on human health: A regression discontinuity analysis in Toronto, Canada. *The Lancet Planetary Health*, 2(1):e19–e26, January 2018. ISSN 25425196. doi: 10.1016/S2542-5196(17)30185-7.

Christensen, Garret and Freese, Jeremy and Miguel, Edward. *Transparent and reproducible social science research*. University of California Press, 2019.

Cinelli, Carlos and Hazlett, Chad. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, 2020.

Open Science Collaboration and others. Estimating the reproducibility of psychological science. *Science*, 349(6251), 2015.

Deryugina, Tatyana and Heutel, Garth and Miller, Nolan H. and Molitor, David and Reif, Julian. The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction. *American Economic Review*, 109(12):4178–4219, December 2019. ISSN 0002-8282. doi: 10.1257/aer.20180279.

Di, Qian and Dai, Lingzhen and Wang, Yun and Zanobetti, Antonella and Choirat, Christine and Schwartz, Joel D. and Dominici, Francesca. Association of Short-term Exposure to Air Pollution With Mortality in Older Adults. *JAMA*, 318(24):2446, December 2017. ISSN 0098-7484. doi: 10.1001/jama.2017.17923.

Dominici, Francesca and Zigler, Corwin. Best Practices for Gauging Evidence of

Causality in Air Pollution Epidemiology. *American Journal of Epidemiology*, 186 (12):1303–1309, December 2017. ISSN 0002-9262, 1476-6256. doi: 10.1093/aje/kwx307.

Dorie, Vincent and Harada, Masataka and Carnegie, Nicole Bohme and Hill, Jennifer. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in medicine*, 35(20):3453–3470, 2016.

Ebenstein, Avraham and Frank, Eyal and Reingewertz, Yaniv. Particulate Matter Concentrations, Sandstorms and Respiratory Hospital Admissions in Israel. 17:6, 2015.

Ferraro, Paul J. and Shukla, Pallavi. Feature—Is a Replicability Crisis on the Horizon for Environmental and Resource Economics? *Review of Environmental Economics and Policy*, 14(2):339–351, June 2020. ISSN 1750-6816. doi: 10.1093/reep/reaa011.

Forastiere, Laura and Carugno, Michele and Baccini, Michela. Assessing short-term impact of PM10 on mortality using a semiparametric generalized propensity score approach. *Environmental Health*, 19(1):46, December 2020. ISSN 1476-069X. doi: 10.1186/s12940-020-00599-6.

Gelman, Andrew and Carlin, John. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, 9(6): 641–651, November 2014. ISSN 1745-6916. doi: 10.1177/1745691614551642.

Gelman, Andrew and Tuerlinckx, Francis. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, 15(3):373–390, September 2000. ISSN 1613-9658. doi: 10.1007/s001800000040.

Gelman, Andrew and Hill, Jennifer and Vehtari, Aki. *Regression and other stories*. Cambridge University Press, 2020.

Giaccherini, Matilde and Kopinska, Joanna and Palma, Alessandro. When particulate matter strikes cities: Social disparities and health costs of air pollution. *Jour-*

nal of Health Economics, page 102478, 2021.

Godzinski, Alexandre and Suarez Castillo, Milena. Short-term health effects of public transport disruptions: air pollution and viral spread channels. 2019.

Greenland, Sander. Invited commentary: The need for cognitive science in methodology. *American journal of epidemiology*, 186(6):639–645, 2017.

Griffin, Beth Ann and Schuler, Megan S and Stuart, Elizabeth A and Patrick, Stephen and McNeer, Elizabeth and Smart, Rosanna and Powell, David and Stein, Bradley and Schell, Terry and Pacula, Rosalie Liccardo. Variation in performance of commonly used statistical methods for estimating effectiveness of state-level opioid policies on opioid-related mortality. Technical report, National Bureau of Economic Research, 2020.

Halliday, Timothy J and Lynham, John and de Paula, Áureo. Vog: Using Volcanic Eruptions to Estimate the Health Costs of Particulates. *The Economic Journal*, 129 (620):1782–1816, May 2019. ISSN 0013-0133, 1468-0297. doi: 10.1111/ecoj.12609.

He, Guojun and Fan, Maoyong and Zhou, Maigeng. The effect of air pollution on mortality in china: Evidence from the 2008 beijing olympic games. *Journal of Environmental Economics and Management*, 79:18–39, 2016.

He, Guojun and Liu, Tong and Zhou, Maigeng. Straw burning, pm2. 5, and death: evidence from china. *Journal of Development Economics*, 145:102468, 2020.

Ioannidis, John P. A. Why Most Discovered True Associations Are Inflated. *Epidemiology*, 19(5):640–648, 2008a.

Ioannidis, John P. A. and Stanley, T. D. and Doucouliagos, Hristos. The Power of Bias in Economics Research. *The Economic Journal*, 127(605):F236–F265, October 2017. ISSN 0013-0133. doi: 10.1111/ecoj.12461.

Ioannidis, John PA. Why most discovered true associations are inflated. *Epidemiology*, pages 640–648, 2008b.

Isphording, Ingo E. and Pestel, Nico. Pandemic meets pollution: Poor air quality increases deaths by COVID-19. *Journal of Environmental Economics and Management*, 108:102448, July 2021. ISSN 00950696. doi: 10.1016/j.jeem.2021.102448.

Jans, Jenny and Johansson, Per and Nilsson, J. Peter. Economic status, air quality, and child health: Evidence from inversion episodes. *Journal of Health Economics*, 61:220–232, September 2018. ISSN 01676296. doi: 10.1016/j.jhealeco.2018.08.002.

Jia, Ruixue and Ku, Hyejin. Is China's Pollution the Culprit for the Choking of South Korea? Evidence from the Asian Dust. *The Economic Journal*, 129(624):3154–3188, November 2019. ISSN 0013-0133, 1468-0297. doi: 10.1093/ej/uez021.

Kim, Sun-young and Sheppard, Lianne and Hannigan, Michael P. and Dutton, Steven J. and Peel, Jennifer L. and Clark, Maggie L. and Vedal, Sverre. The sensitivity of health effect estimates from time-series studies to fine particulate matter component sampling schedule. *Journal of Exposure Science and Environmental Epidemiology; Tuxedo*, 23(5):481–6, September 2013. ISSN 15590631. doi: http://dx.doi.org.ezproxy.cul.columbia.edu/10.1038/jes.2013.28.

Knittel, Christopher R. and Miller, Douglas L. and Sanders, Nicholas J. Caution, Drivers! Children Present: Traffic, Pollution, and Infant Health. *Review of Economics and Statistics*, 98(2):350–366, May 2016. ISSN 0034-6535, 1530-9142. doi: 10.1162/REST_a_00548.

Le Tertre, A and Medina, S and Samoli, E and Forsberg, B and Michelozzi, P and Boumghar, A and Vonk, JM and Bellini, A and Atkinson, R and Ayres, JG and others. Short-term effects of particulate air pollution on cardiovascular diseases in eight european cities. *Journal of Epidemiology & Community Health*, 56(10): 773–779, 2002.

Liu, Cong and Chen, Renjie and Sera, Francesco and Vicedo-Cabrera, Ana M. and Guo, Yuming and Tong, Shilu and Coelho, Micheline S.Z.S. and Saldiva, Paulo

H.N. and Lavigne, Eric and Matus, Patricia and Valdes Ortega, Nicolas and Osorio Garcia, Samuel and Pascal, Mathilde and Stafoggia, Massimo and Scortichini, Matteo and Hashizume, Masahiro and Honda, Yasushi and Hurtado-Díaz, Magali and Cruz, Julio and Nunes, Baltazar and Teixeira, João P. and Kim, Ho and Tobias, Aurelio and Íñiguez, Carmen and Forsberg, Bertil and Åström, Christofer and Ragettli, Martina S. and Guo, Yue-Leon and Chen, Bing-Yu and Bell, Michelle L. and Wright, Caradee Y. and Scovronick, Noah and Garland, Rebecca M. and Milojevic, Ai and Kyselý, Jan and Urban, Aleš and Orru, Hans and Indermitte, Ene and Jaakkola, Jouni J.K. and Ryti, Niilo R.I. and Katsouyanni, Klea and Analitis, Antonis and Zanobetti, Antonella and Schwartz, Joel and Chen, Jianmin and Wu, Tangchun and Cohen, Aaron and Gasparrini, Antonio and Kan, Haidong. Ambient Particulate Air Pollution and Daily Mortality in 652 Cities. *New England Journal of Medicine*, 381(8):705–715, August 2019. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa1817364.

Liu, Yan and Yan, Zhijun and Liu, Su and Wu, Yuting and Gan, Qingmei and Dong, Chao. The effect of the driving restriction policy on public health in Beijing. *Natural Hazards*, 85(2):751–762, January 2017. ISSN 0921-030X, 1573-0840. doi: 10.1007/s11069-016-2602-8.

Mayer, Michael. missRanger: Fast Imputation of Missing Values. Comprehensive R Archive Network (CRAN), 2019.

McShane, Blakeley B and Gal, David and Gelman, Andrew and Robert, Christian and Tackett, Jennifer L. Abandon statistical significance. *The American Statistician*, 73(sup1):235–245, 2019.

Moretti, Enrico and Neidell, Matthew. Pollution, Health, and Avoidance Behavior: Evidence from the Ports of Los Angeles. *Journal of Human Resources*, 46(1):154–175, 2011. ISSN 1548-8004. doi: 10.1353/jhr.2011.0012.

Mullins, Jamie and Bharadwaj, Prashant. Effects of Short-Term Measures to Curb

Air Pollution: Evidence from Santiago, Chile. *American Journal of Agricultural Economics*, 97(4):1107–1134, July 2015. ISSN 0002-9092, 1467-8276. doi: 10.1093/ajae/aau081.

Orellano, Pablo and Reynoso, Julieta and Quaranta, Nancy and Bardach, Ariel and Ciapponi, Agustin. Short-term exposure to particulate matter (pm10 and pm2.5), nitrogen dioxide (no2), and ozone (o3) and all-cause and cause-specific mortality: Systematic review and meta-analysis. *Environment international*, 142:105876, 2020.

Peng, Roger D and Dominici, Francesca. Statistical methods for environmental epidemiology with r. *R: a case study in air pollution and health*, 2008.

Peng, Roger D and Dominici, Francesca and Louis, Thomas A. Model choice in time series studies of air pollution and mortality. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(2):179–203, 2006.

Romer, David. In praise of confidence intervals. In *AEA Papers and Proceedings*, volume 110, pages 55–60, 2020.

Rosenbaum, Paul R. *Design of observational studies*, volume 10. Springer, 2010.

Rubin, Donald B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. ISSN 0022-0663. doi: 10.1037/h0037350.

Samet, Jonathan M and Zeger, Scott L and Dominici, Francesca and Curriero, Frank and Coursac, Ivan and Dockery, Douglas W and Schwartz, Joel and Zanobetti, Antonella. The national morbidity, mortality, and air pollution study. *Part II: morbidity and mortality from air pollution in the United States Res Rep Health Eff Inst*, 94(pt 2):5–79, 2000.

Schäfer, Thomas and Schwarz, Marcus A. The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, 10:813, 2019.

Schell, Terry L and Griffin, Beth Ann and Morral, Andrew R. *Evaluating methods to estimate the effect of state laws on firearm deaths: A simulation study*. Rand, 2018.

Schlenker, Wolfram and Walker, W. Reed. Airports, Air Pollution, and Contemporaneous Health. *The Review of Economic Studies*, 83(2):768–809, April 2016. ISSN 0034-6527, 1467-937X. doi: 10.1093/restud/rdv043.

Schwartz, Joel. What are people dying of on high air pollution days? *Environmental research*, 64(1):26–35, 1994.

Schwartz, Joel and Austin, Elena and Bind, Marie-Abele and Zanobetti, Antonella and Koutrakis, Petros. Estimating Causal Associations of Fine Particles With Daily Deaths in Boston: Table 1. *American Journal of Epidemiology*, 182(7):644–650, October 2015. ISSN 0002-9262, 1476-6256. doi: 10.1093/aje/kwv101.

Schwartz, Joel and Bind, Marie-Abele and Koutrakis, Petros. Estimating Causal Effects of Local Air Pollution on Daily Deaths: Effect of Low Levels. *Environmental Health Perspectives*, 125(1):23–29, January 2017. ISSN 0091-6765, 1552-9924. doi: 10.1289/EHP232.

Schwartz, Joel and Fong, Kelvin and Zanobetti, Antonella. A National Multicity Analysis of the Causal Effect of Local Pollution, NO2, and PM2.5 on Mortality. *Environmental Health Perspectives*, 126(8):087004, August 2018. ISSN 0091-6765, 1552-9924. doi: 10.1289/EHP2732.

Shah, Anoop SV and Lee, Kuan Ken and McAllister, David A and Hunter, Amanda and Nair, Harish and Whiteley, William and Langrish, Jeremy P and Newby, David E and Mills, Nicholas L. Short term exposure to air pollution and stroke: systematic review and meta-analysis. *bmj*, 350, 2015.

Sheldon, Tamara L. and Sankaran, Chandini. The Impact of Indonesian Forest Fires on Singaporean Pollution and Health. *American Economic Review*, 107(5):526–529, May 2017. ISSN 0002-8282. doi: 10.1257/aer.p20171134.

Simonsohn, Uri and Nelson, Leif D and Simmons, Joseph P. P-curve: a key to the

file-drawer. *Journal of experimental psychology: General*, 143(2):534, 2014.

Smaldino, Paul E and McElreath, Richard. The natural selection of bad science. *Royal Society open science*, 3(9):160384, 2016.

Drew Stommes and P. M. Aronow and Fredrik Sävje. On the reliability of published findings using the regression discontinuity design in political science, 2021.

Timm, Andrew. Retrodesign: Tools for Type S (Sign) and Type M (Magnitude) Errors. Comprehensive R Archive Network (CRAN), March 2019.

VanderWeele, Tyler J and Ding, Peng. Sensitivity analysis in observational research: introducing the e-value. *Annals of internal medicine*, 167(4):268–274, 2017.

Vasishth, Shravan and Gelman, Andrew. How to embrace variation and accept uncertainty in linguistic and psycholinguistic data, 2019.

Vichit-Vadakan, Nuntavarn and Vajanapoom, Nitaya and Ostro, Bart. The Public Health and Air Pollution in Asia (PAPA) Project: Estimating the mortality effects of particulate matter in Bangkok, Thailand. *Environmental Health Perspectives*, 116(9):1179–1182, September 2008. ISSN 0091-6765. doi: 10.1289/ehp.10849.

Wasserstein, Ronald L and Lazar, Nicole A. The asa statement on p-values: context, process, and purpose, 2016.

Williams, Austin M. and Phaneuf, Daniel J. and Barrett, Meredith A. and Su, Jason G. Short-term impact of $PM_{2.5}$ on contemporaneous asthma medication use: Behavior and the value of pollution reductions. *Proceedings of the National Academy of Sciences*, 116(12):5246–5253, March 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1805647115.

Winquist, Andrea and Klein, Mitchel and Tolbert, Paige and Sarnat, Stefanie Ebelt. Power estimation using simulations for air pollution time-series studies. *Environmental Health*, 11(1):1–12, 2012.

Young, Alwyn. Consistency without inference: Instrumental variables in practical application. 2019.

Zhong, Nan and Cao, Jing and Wang, Yuzhu. Traffic congestion, ambient air pollution, and health: Evidence from driving restrictions in beijing. *Journal of the Association of Environmental and Resource Economists*, 4(3):821–856, 2017.

Ziliak, Stephen Thomas and McCloskey, Deirdre N. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Economics, Cognition, and Society. University of Michigan Press, Ann Arbor, 2008a. ISBN 978-0-472-07007-7 978-0-472-05007-9.

Ziliak, Steve and McCloskey, Deirdre Nansen. *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. University of Michigan Press, 2008b.