Inference Design In Studies on Acute Health Effects of Air Pollution

Vincent Bagilet[1]

Léo Zabrocki[2]

July 7, 2021

Abstract

We explore statistical power characteristics of various empirical strategies implemented to estimate the short-term health effect of air pollution. Through an extensive literature review, we retrieve the estimates and standard errors of a large number of studies published on this topic. We find that a non-negligible share of studies may suffer from low power issues and could thereby exaggerate effect sizes. The analysis of published results highlights potential shortcomings of the literature but does not enable to precisely identify drivers of theses issues. We therefore run realistic simulations to investigate how statistical power varies with the treatment effect size, the number of observations, the proportion of treated units as well as the distribution of the outcome. Usual causal identification methods implemented in this literature, such as instrumental variable (IV), regression discontinuity design (RD) or difference-in-differences (DiD), may yield overestimated effect sizes. This issue is driven by the imprecision of the IV estimator and the small number of exogenous shocks usually exploited in DiD and RD designs. When focusing on particular groups such as the elderly or children, researchers should be aware that statistical power is lowered by the limited average count of health outcomes.

[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

In the last decade, researchers in economics and epidemiology have made considerable efforts to credibility estimate the acute health effects of air pollution. Research designs based on causal inference methods have helped them better address the issue of unmeasured confounding variables (Dominici and Zigler 2017, Bind 2019). Newly obtained results have direct policy implications as they often strengthen the case for lowering thresholds of air quality alerts (Schwartz et al. 2015; 2018, Deryugina et al. 2019). While considerations on identification strategies are pivotal to these papers, statistical inference issues are rarely discussed. Working with observational data within the null hypothesis significance testing framework does not invite to pay attention to statistical power. Yet, estimates that are deemed statistically significant tend to overestimate true effect sizes when the statistical power is low (Ioannidis 2005, Gelman and Carlin 2014, Ioannidis et al. 2017, Altoè et al. 2020, van Buuren and Greenacre). This issue is 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.

In this paper, we undertake the first empirical evaluation of the *inference design* of studies on the short-term health effects of air pollution. By "evaluating an inference design", we mean to assess whether the statistical power of these observational studies could lead to overestimated acute effects of air pollution on health outcomes.

First, we carry out a *retrospective* analysis of studies investigating the short-term effects of air pollution on mortality or morbidity. With an extensive search strategy, we retrieve most articles published in the standard epidemiology literature and all articles that we are aware of based on causal inference methods. We follow the approach proposed by Gelman and Carlin (2014) to compute, based on hypothetical true effect sizes, the statistical power and the exaggeration factor of statistically significant estimates. We then develop a *prospective* design analysis to help researchers evaluate potential inference issues associated with their identification strategies. We run simulations specific to each causal inference method, based on real data from the US National Morbidity, Mortality, and Air Pollution Study. We study how inference issues are affected by the treatment effect size, the number of observations, the proportion of treated units, and the distribution of the health outcome.

While many studies do not suffer from power issues, we find that a substantial share does and could greatly overestimates the short term health effects of air pollution. The quality of inference designs does not appear to be a central issue in

this field as the proportion of studies with low power has remained high over time. In the causal inference literature, we observe a clear negative linear correlation between effect sizes and precision of estimates. Less precise studies may discover true causal association but their estimates are likely to be inflated.

Our simulation results complete our systematic analysis of the literature. First, we find that instrumental variable estimates are more likely to overestimate true effect sizes than naive estimates. This issue is even present for very large first stage *F*-statistics. Second, many papers in the literature exploit rare events such as public transport strikes, thermal inversions, or air quality alerts as exogenous shocks on air pollution. Settings with few treated units have a dramatically low power and greatly exaggerate true effect sizes. There is therefore a trade-off between the use of such sporadic events to obtain unbiased estimates and the risk of overestimating true treatment effects as a result of low power issues. Third, many researchers have investigated acute effects of air pollution on health outcomes with small daily average counts. We find that with very few daily cases of an health outcome, statistical power is extremely low, even for large sample sizes. In such circumstances, researchers have very high chances to overestimate the effects of air pollution.

Our paper is organized as follows. In Section II, we implement a simple simulation exercise to help readers understand why a statistically significant estimate exaggerate the true effect when statistical power is low. In section III, we discuss the results of our retrospective analysis of the literature. In Section IV, we detail how we build our simulations. We display the results of these simulations in section V and we provide specific guidance on study design for researchers in Section VI.

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 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 city over a one year period. 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 $PM_{2.5}$ of a European capital city.

Table 1: Science Table of the Experiment.

Day Index	$Y_i(0)$	$Y_i(1)$	W_i	Y_i^{obs}
1	123	124	1	124
2	79	80	1	80
3	83	84	0	83
⋮	⋮	⋮	⋮	⋮
363	136	137	1	137
364	106	107	0	106
365	95	96	0	95

Notes: This table displays the potential outcomes, the treatment status and the observed outcomes for 6 of the 365 daily units in the scientist's experiment.

To simulate this experiment, we create a Science table where we observe the pair of potential outcomes of each day, $Y_i(W_i = 0)$ and $Y_i(W_i = 1)$ (see Table 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 create the $Y_i(0)$, i.e., the daily non-accidental mortality counts in the absence of treatment by drawing 365 observations from a Negative Binomial distribution with a mean of 106 and a variance of 402. We choose the parameters to approximate the distribution of non-accidental mortality counts in a large European city. We then consider a constant treatment effect of 1 additional death due to the air pollution increase such that $Y(1)_i = Y(0)_i + 1$. It represents approximately a 1% increase in the mean of the outcome or a 0.05 increase in its standard deviation. 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 10µg/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.

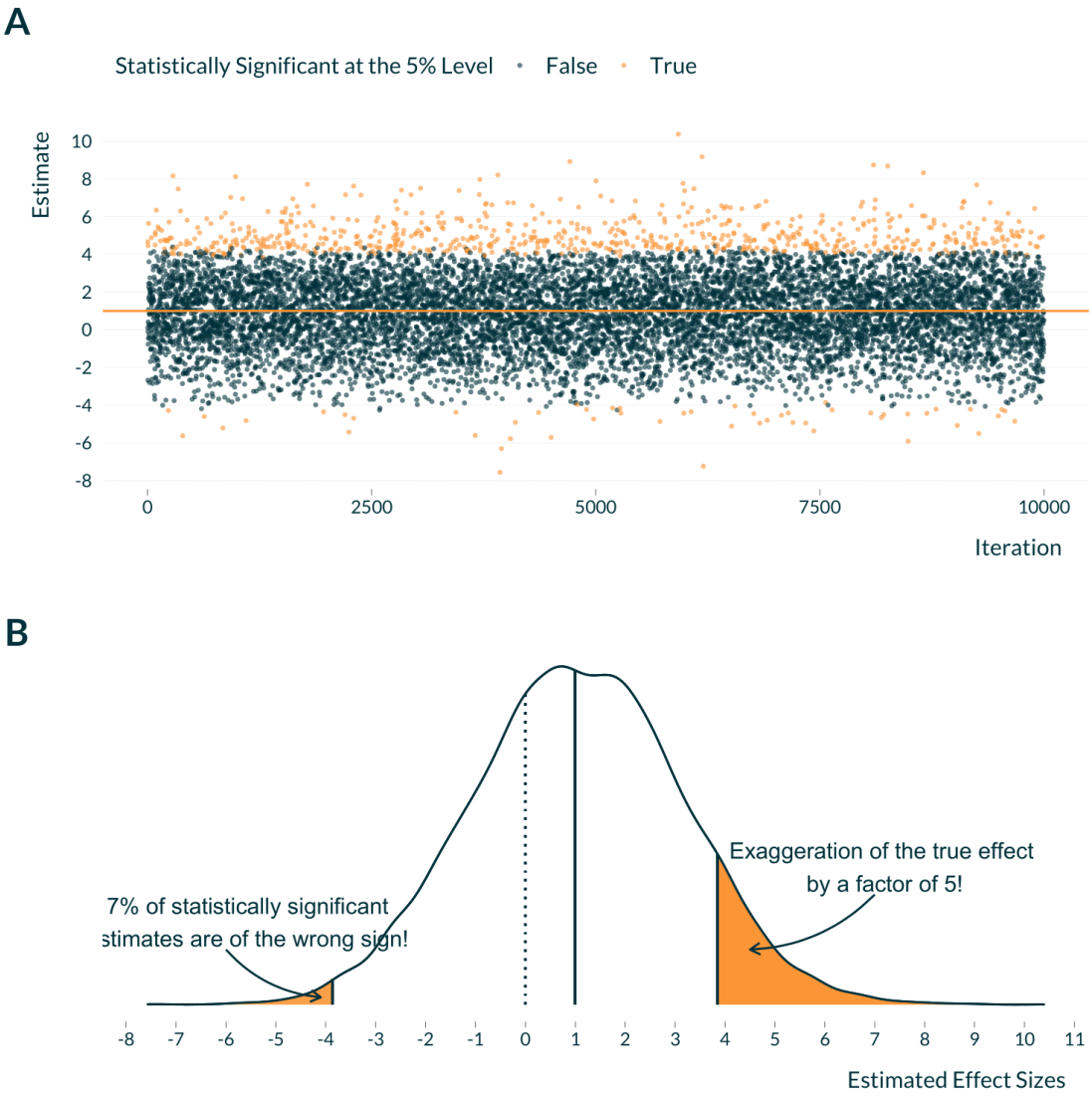
The scientist implements a complete experiment where they randomly allocate half of the days to the treatment group and the other half to the control group. Following the fundamental problem of causal inference, the daily count of deaths they observe is given by the equation: $Y_i^{obs} = W_i \times Y_i(1) + (1 - W_i) \times Y_i(0)$. Treated units express their $Y_i(1)$ values and control units their $Y_i(0)$ values. The scientist computes the average difference in means between treated and control outcomes and obtains an estimate for the treatment effect of 4 additional deaths, with a p-value of $\simeq 0.04$. The estimate is "statistically significant" at the 5% level. The "significant" result fulfills the scientist expectations, who immediately starts writing their paper. Had they not obtained a statistically significant estimate, they would have not carried on with the whole publication process.

Unfortunately for the scientist, we are in a position where we have much more information than him. We observe the two potential outcomes for each day and know that the treatment effect is equal to +1 daily death. To gauge the inference properties of an experiment with a sample size of 365 days, we replicate this experiment 10,000 times.

2.2 Defining Statistical Power, Type M and S errors

In Figure 1, we plot the estimates of the 10,000 iterations of the experiments (Panel A) and their density distribution (Panel B). The average of estimates is equal to 1 additional death, the true effect size. The statistical power is the probability to get a significant estimate when there is actually an effect. For this experiment, it can be computed as the proportion of estimates that are statistically significant estimates at the 5% level. Among the 10,000 iterations, only 700 estimates are statistically significant (the orange dots in Panel A): the statistical power of the scientist for the experiment is therefore 7%. The scientist was therefore "lucky" to get a statistically significant estimate. We then evaluate the extent to which statistically significant estimates exaggerate the true effect of $PM_{2.5}$ on mortality. The exaggeration ratio, also called the type M error by Gelman and Carlin (2014), is computed as the average of the ratio of the absolute values of the statistically significant estimates over the true effect size. If the scientist happens to get a statistically significant estimate, it would, in expectation, overestimate the true effect size by a factor of 5! Strikingly, a fraction of statistically significant estimates are of the wrong sign in Figure 1. This leads to the definition of the type S error as the probability that the estimate, when statistically significant, has a sign opposed to the true effect (Gelman and Tuerlinckx

Figure 1: Estimates of the 10,000 Simulations.



Notes: In Panel A, blue and orange dots represent the point estimates of the 10,000 iterations of the randomized experiment ran by the mad scientist. Orange dots are statistically significant at the 5% level while blue dots are not. The orange solid line represents the true effect. Panel B displays the density distribution of the 10,000 estimates of the treatment effect. The solid blue line is the average of the estimates and is equal to the true effect, +1. The orange areas represent the proportion of statistically significant estimates at the 5% level.

2000). For this experiment, a statistically significant estimate has a 7% probability of being of the wrong sign. The type M error and the probability to make a type S error are high for this experiment due to the particularly small sample and true effect sizes. With a larger sample size, the statistical power would rise and conversely

type M and S error would shrink.

2.3 Relevance for Studies on Acute Health Effects of Air Pollution

If the scientist could replicate many times his experiment, they would find on average the correct effect size of a 10 $\mu g/m^3$ increase in $PM_{2.5}$ on daily mortality. Septic readers could rightly wonder why they should worry about type M and S errors. Researchers are—despite recent changes in scientific practices—not incited enough to publish replication exercises and non statistically significant estimates. As a consequence, published estimates being mostly selected among statistically significant ones may overestimate true effect size. Type M and S errors highlight the arguably counter-intuitive danger of having too much confidence in statistically significant estimates when studies are under-powered. These concepts are highly relevant for estimating the acute health effects of air pollution as signal to noise ratios are typically low in this literature. Effect sizes are often remarkably small and modeling the variations in health outcome counts to reduce noise is especially challenging (Black et al. 2019). Large sample sizes are also required to precisely estimate the acute effect of air pollution.

3 Retrospective Analysis of the Literature

In this section, we assess whether the standard and causal inference literature suffer from statistical power issues. Beforehand, we describe the procedure to run a retrospective design analysis for a study.

3.1 Computing Statistical Power, Type M and S Errors

Running a retrospective design analysis for a study, i.e., computing its statistical power, type M and type S errors, 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 an article's research design, such as the number of observations, are assumed to be fixed. The R package `retrodesign` developed by Timm (2019) implements the closed-form expressions derived by Lu et al. (2019) making it very easy to carry out the procedure for a given study. Thinking about the true effect size one is trying to estimate is the central piece of a retrodesign analysis. As the true effect is never observed, researchers can have very different priors on its magnitude.

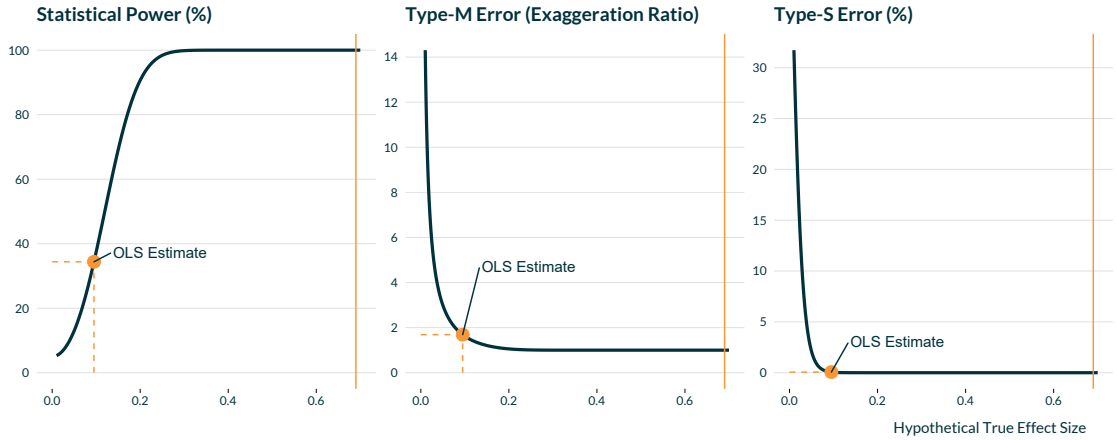
They could therefore assess differently the extent to which a study risks to suffer from statistical power issues. We find two different ways to think about true effect sizes.

First, we take a comprehensive approach to evaluate the quality of the inference design of studies in the literature. To be informative, a study needs to have enough statistical power to precisely estimate a range of effect sizes that one would deem credible. Hence, a well-designed study should be able to detect effects even though they are smaller than the estimated one. If assuming that the true effect is 3/4 of the measured effect only yields a power of 30%, the study is certainly underpowered. If this analysis could be replicated, with this design, such an effect would not be detected in 70% of the replications. For each study, we thus compute power, type M and S errors assuming that the true effect size is equal to a fraction of the estimated one. This exercise constitutes a broad sensitivity analysis. Second, we can try to guess, for each study, what could be the value of the true effect size. We do such an analysis only for the causal inference literature as we are able to carefully read each paper. For each causal inference paper, we try to find the most similar paper that uses non-causal method and record what would be the effect sizes predicted [still need to be done]. In many cases, we find that causal estimates were an order of magnitude higher than what the standard literature had previously found—this could be explained by the fact that causal inference methods better overcome omitted variables bias and in some cases measurement error issues. Yet, it would be very interesting to know if causal inference designs have enough statistical power to precisely estimate the effect sizes found intend the standard literature. We also run design calculations for instrumental variable strategies assuming that the true effect size is equal to the estimate of the associated standard multivariate model. Instrumental variable estimates are well-know to be less precise and could be higher than standard multivariate estimates just because they run into a type M error.

Finally, as Type M and S errors are new concepts to most researchers, we provide a case study to illustrate how a retrospective analysis can be concretely carried out. 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, "a 1 $\mu g/m^3$ (about 10 percent of the mean) increase in $PM_{2.5}$ exposure for one day causes [0.69 $\pm$ 0.061] additional

deaths per million elderly individuals over the three-day window that spans the day of the increase and the following two days". In Figure 2, we plot power, type M and S errors as a function of hypothetical true effect sizes. The solid orange line represents the observed two-stage least square estimate reported in the article.

Figure 2: Power, Type M and S Errors Curves for Deryugina et al. (2019)



Notes: In each panel, a metrics, such as the statistical power, the exaggeration ratio or the probability to make a type S error, is plotted against a range of hypothetical effect sizes. The solid orange line represents the observed two-stage least square estimate reported in the article.

The estimate found by Deryugina et al. (2019) represents a relative increase of 0.18% in mortality. Is this estimated effect size large compared to those reported in other articles? We found two similar articles to draw a comparison. Using a case-crossover design and conditional logistic regression, Qian Di (citation) find that a one $\mu g/m^3$ 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. Schwartz et al. (2018) estimate that a one $\mu g/m^3$ increase in $PM_{2.5}$ instrumented concentrations with the planetary boundary layer, wind speed, and air pressure leads to a 0.15% increase in non-accidental mortality. The effect size found by Deryugina et al. (2019) is a bit higher but relatively close to these two studies. Given the sample size and the similarity of the estimated effect compared to other studies, Deryugina et al. (2019) have likely a high statistical power and are unlikely to make a type M error [add vertical lines on the graph for these estimates]. 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 overestimation factor would be 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 "naive" estimate. Yet, the authors could argue that their instrumental variable strategy leads to a higher effect size as it overcome 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.

3.2 Causal Inference Literature

Using Google Scholar, PubMed, and journal websites, 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 studying 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.

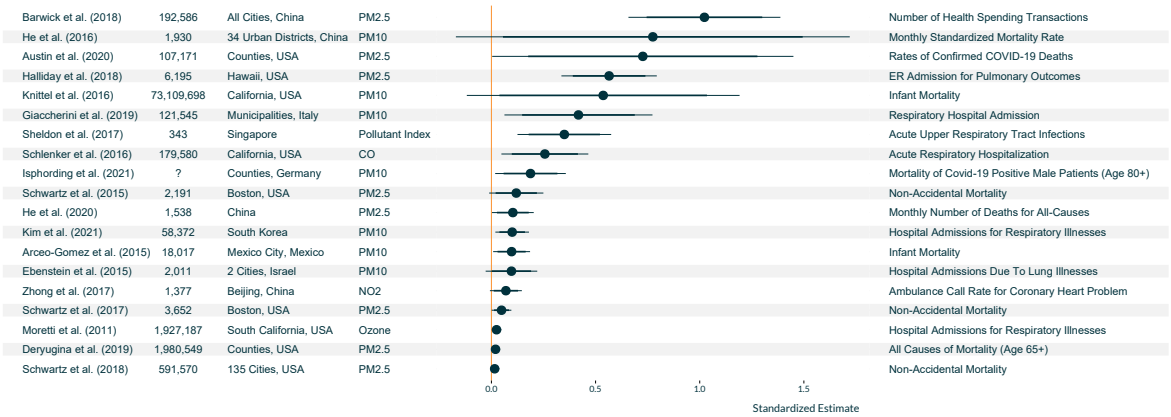
Table 2: Our Corpus of Papers from the Causal Inference Literature.

Article	Location	Health Outcome	Independent Variables	Study Design
Arceo et al. (2016)	Mexico City, Mexico	Infant Mortality	PM10, Thermal Inversion (IV)	Instrumental Variable
Austin 2020	Counties, USA	Rates of Confirmed COVID-19 Deaths	PM2.5 (air pollutant), Wind Direction (IV)	Instrumental Variable
Baccini 2017	Milan, Italy	Non-Accidental Mortality	Dummy for PM10 Concentration >To 40 $\mu g/m^3$	Propensity Score Matching
Barwick 2018	All Cities, China	Number of Health Spending Transactions	PM2.5, Spatial Spillovers of PM2.5 (IV)	Instrumental Variable
Bauernschuster 2017	5 Largest Cities, Germany	Admissions for Abnormalities of Breathing (age below 5)	PM10, Public Transport Strikes Dummy	Difference in Differences
Beard 2012	Salt Lake County, USA	Emergency Visits For Asthma	Thermal Inversions	Time-stratified case-crossover design
Chen 2018	Toronto, Canada	Asthma-Related Emergency Department Visits	Air Quality Eligibility, Air Quality Altert	Fuzzy Regression Discontinuity
Deryugina 2019	Counties, USA	All Causes of Mortality (Age 65+)	PM2.5, Wind Direction (IV)	Instrumental Variable
Ebenstein 2015	2 Cities, Israel	Hospital Admissions Due To Lung Illnesses	PM10 (air pollutant), Sandstorms (IV)	Instrumental Variable
Forastiere 2020	Milan, Italy	Non-Accidental Mortality	Setting PM10 Daily Exposure Levels >To 40 $\mu g/m^3$ To 40	Generalized Propensity Score
Giaccherini 2019	Municipalities, Italy	Respiratory Hospital Admission	PM10, Public Transport Strikes	Difference in Differences
Godzinski 2019	10 Cities, France	Emergency Admissions for Upper Respiratory System (Age 0-4)	CO, Public Transport Strikes	Difference in Differences
Halliday 2018	Hawaii, USA	ER Admission for Pulmonary Outcomes	PM2.5, SO2 Emissions From Kilauea Volcano and Wind Direction (IV)	Instrumental Variable
He 2016	34 Urban Districts, China	Monthly Standardized Mortality Rate	PM10, Regulation and Traffic Control Status (IV)	Instrumental Variable
He 2020	China	Monthly Number of Deaths for All-Causes	PM2.5, Straw Burning (IV)	Instrumental Variable
Isphording 2021	Counties, Germany	Mortality of Covid-19 Positive Male Patients (Age 80+)	PM10, Wind direction (IV)	Instrumental Variable
Jans 2018	Sweden	Children Health Care Visits for Respiratory Illness	PM10, Thermal Inversion (IV)	Instrumental Variable
Jia 2019	South Korea	Mortality Rates for Respiratory and Cardiovascular Diseases	Dusty Days Times China's AQI	Reduced-Form
Kim 2021	South Korea	Hospital Admissions for Respiratory Illnesses	PM10 (air pollutant), Average PM10 Level By Date (IV)	Instrumental Variable
Knittel 2016	California, USA	Infant Mortality	PM10, Road Traffic Flow and Weather variables (IV)	Instrumental Variable
Moretti 2011	South California, USA	Hospital Admissions for Respiratory Illnesses	O3, Vessel Traffic (IV)	Instrumental Variable
Mullins 2014	Santiago Metropole, Chile	Cumulative Deaths (age >64)	PM10, Air quality Alerts	Matching + Difference in Differences
Schlenker 2016	California, USA	Acute Respiratory Hospitalization	CO, Planes Taxi Time (IV)	Instrumental Variable
Schwartz 2015	Boston, USA	Non-Accidental Mortality	PM2.5, Back Trajectories of PM2.5 (IV)	Instrumental Variable
Schwartz 2017	Boston, USA	Non-Accidental Mortality	PM2.5, Height Of Planetary Boundary Layer and Wind Speed (IV)	Instrumental Variable
Schwartz 2018	135 Cities, USA	Non-Accidental Mortality	PM2.5, Planetary Boundary Layer, Wind Speed, and Air Pressure (IV)	Instrumental Variable
Scheldon 2017	Singapore	Acute Upper Respiratory Tract Infections	Pollutant Index, Indonesian Fire Radiative Power (IV)	Instrumental Variable
Williams 2018	USA	Asthma Rescue Event	PM2.5	Poisson fixed-effects models
Zhong 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.

In Figure 3, we display the estimated standardized effect sizes and their associated 95% (thick lines) and 99% (thin lines) confidence intervals for articles using an instrumental variable strategy.

Figure 3: Forest Plot of Standardized Estimates Effect Sizes of Articles Based on Instrumental Variable Strategy.

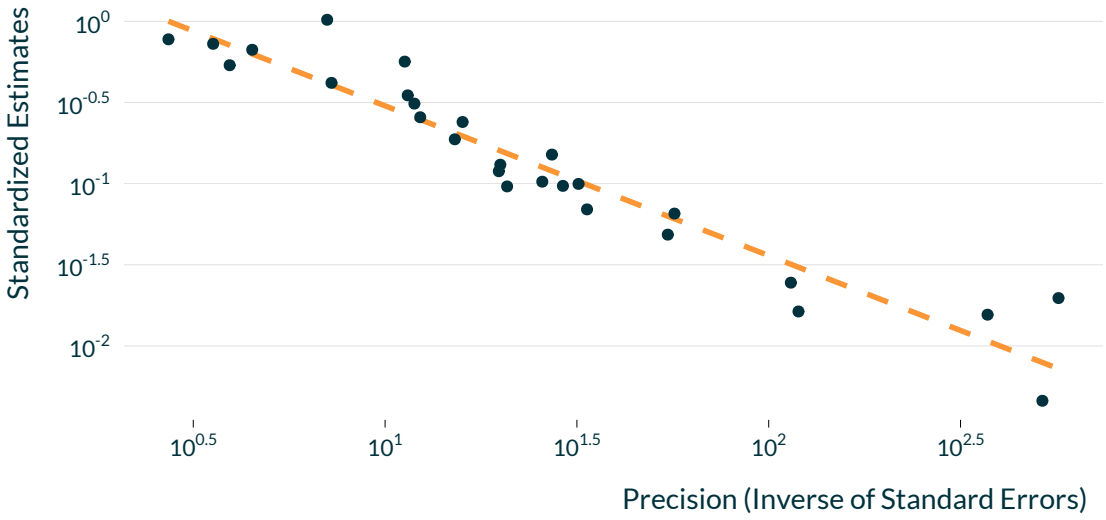


Notes: The first column displays the first author of the article. The second, third and fourth columns reports the sample size, the context of the article and the air pollutant instrument respectively. In the fifth column, each blue dot represents the standardized effect size of an article, the thick lines are the associated 95% confidence intervals and the thine lines are the 99% confidence intervals. In the sixth column, the health outcomes investigated are displayed. [Add columns names with Adobe Illustrator]

Standardized estimated effect sizes vary a lot, from 0.02 to 1, and several studies seem imprecise given their wide confidence intervals. One explanation for this variety of effect sizes is that researchers use different causal methods, rely on different natural experiments and look at different health outcomes. While we think that this may be part of the story, Figure 4 points towards another possible explanation.

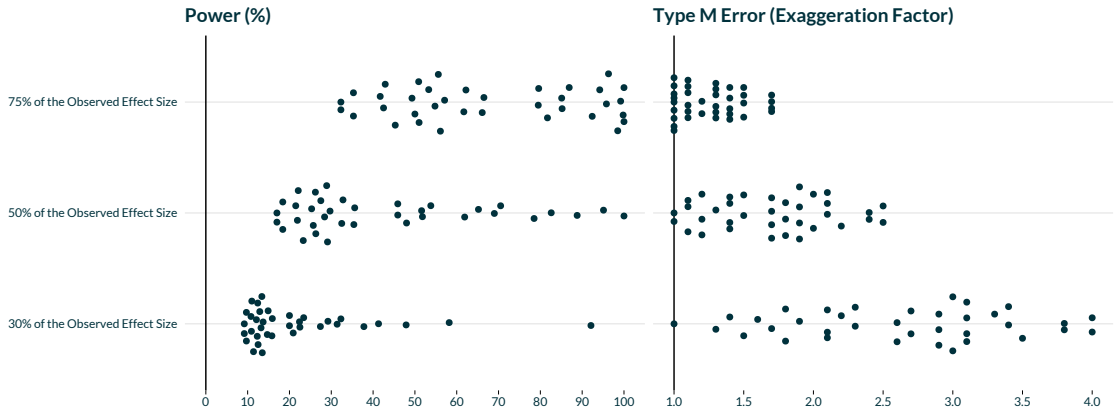
The more precise studies are the ones with the lowest effect sizes. This pattern has been observed in other fields, where replication exercises with larger sample sizes have resulted in deflated effect sizes. We think that this figure could indicate that studies with large effect sizes run into type M error. Alternatively, since public policies are driven by findings from the academic literature, researchers may be incentivized to elicit smaller and smaller effects, therefore needing more and more precise estimates. To evaluate potential inference issues in this literature, we compute the statistical power, the exaggeration factor (type M error) and the probability to get an estimate of the wrong sign (type S error) for all studies based on hypothetical true effect sizes equal to 75%, 50% and 30% of the estimates. Results for the different scenarios are displayed in Figure 5.

Figure 4: Standardized Estimates Against Precision.



Notes: Standardized Estimates 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.

Figure 5: Statistical Power and Type M Error of Causal Inference Studies.

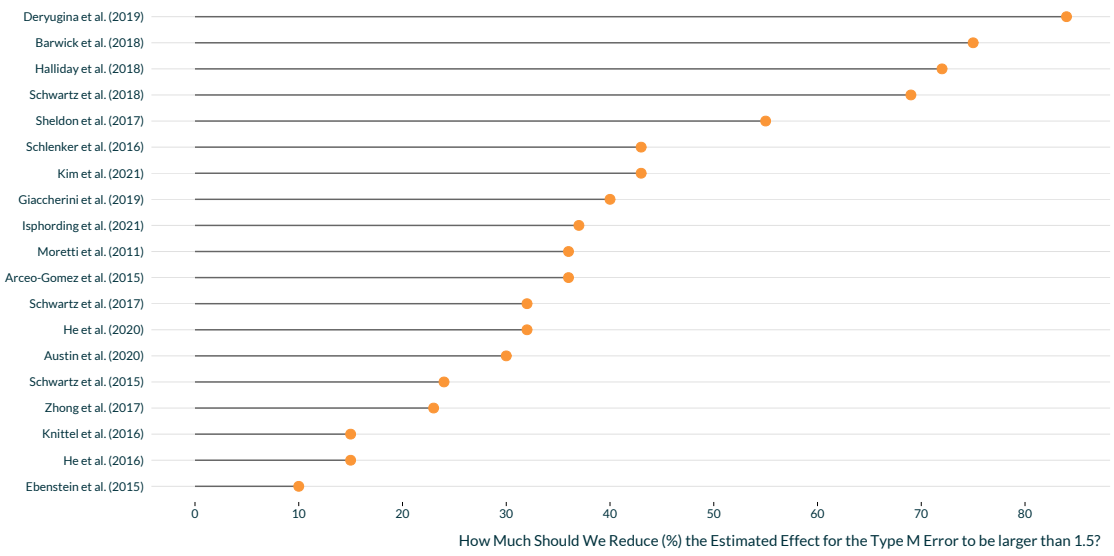


Notes: Each blue dot represents either the statistical power or the exaggeration factor of a study under three hypotheses about its true effect size.

If the true effect size of each study was equal to 75% of the estimate, the average statistical power would be equal to 67% and the Type M error would be 1.3. The causal inference literature could then be relatively confident in the magnitude of its estimated effects. However, if the true effect size of each study was equal to 50% of the estimate, the average statistical power would be 43% and the exaggeration factor

would be equal to 1.7. Finally, in the most extreme scenario where the true effect size of each study was equal to 30% of the estimate, the average statistical power would be 22% and the exaggeration factor would be equal to 2.6. In all scenarios, the probability to make a type S error is nearly null. In Figure 5, we can see that there is a wide heterogeneity in the robustness of studies to inference issues—some of them seem relatively well powered while other seem to 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 2 or 3 times smaller than the ones they find. To illustrate this heterogeneity, we compute by how much the estimated effect size should be decreased for a study to result into a type M error of 1.5. Figure 6, we display the relevant graph for studies based on instrumental variable strategies. For the first 5 studies at the bottom of the graph, if the true effect sizes were 30% inferior to the estimated effect sizes, they would make a type M error of at least 1.5. The quality of the inference for these studies is therefore worrisome. There are however studies that seem to be well-powered as the true effect size would need to be over 70% below the estimated effect in order for them to run into a large type M error.

Figure 6: When Would Studies Exaggerate True Effect Sizes by a Factor of 1.5?



Notes: Orange dots represent the percentage decrease of the estimated effect size needed for a study to make a 1.5 Type M error.

Overall, our exploration of the causal inference literature reveals that it is likely that some studies are under-powered and could run into type M error. It may partly

explain why there is a large heterogeneity in effect sizes in this literature.

3.3 Standard Literature

Contrary to the burgeoning causal inference literature, more than a thousand of papers have been published on the short term health effects of air pollution in epidemiological, medical and public health journals. Most studies rely on generalized additive models 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, they could suffer less from power issues.

We use a search query to retrieve 1834 relevant articles from PubMed and Scopus. We then develop a detection algorithm taking advantage of a standardized reporting procedure of results. Articles in this literature commonly display estimates and confidence intervals in their abstracts, allowing us to extract them using REGular EXpressions (regex). We illustrate the 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 µg/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 valid estimates from 784 articles. The set of articles considered is therefore limited to articles displaying confidence intervals and point estimates in their abstracts. We also build regex queries to retrieve other information about the articles such as the air pollutant and health outcome studied, the length of the study and the number of cities considered.

Contrary to the causal inference literature for which we read each article, we do not know what is exactly measured in each analysis since there is no standardized way of reporting the results beyond mentioning confidence intervals. For instance, studies can express estimated effect sizes either in terms of relative increase in the average daily count of an health outcome or with relative risk. Besides, studies do

not express their estimated effect sizes for a same increase in a pollutant concentration. Running design calculations therefore require us to only compute sensitivity analyses, expressing true effect sizes as a fraction of estimated coefficients.

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 93% and the median exaggeration factor would be nearly 1. At least 50% of this literature does not seem to suffer from substantial power issues. Type S error does not appear to be an important 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 present a substantial risk of making a type M error. About 40% 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 is 1.5. These figures however hide a lot of heterogeneity across studies, which we try to apprehend.

We find that the health outcome and the air pollutant studied do not seem to be related to inference issues. Health science journals appear to be more prone to power issues than other journals, but a shared lack of concern for low power issues is seen across all fields. Researchers appear to be aware that they should work with large sample size as they often carry out multi-city studies and sometimes explicitly state that they investigate non-accidental mortality causes as the average daily count is higher to increase statistical power. Yet, the proportion of low power studies has been stagnating since the 1990s, revealing that practices regarding statistical power have not evolved. Even more worryingly, we find that in recent years, more and more articles display very large type M errors. As in the causal inference literature, the quality of inference designs does not seem to be a central issue for most researchers in this field. This might be explained by the scarcity of guidance on the determinants of statistical power.

4 Prospective Analysis of Causal Inference Methods

The review of the standard and causal literature enables us to get a sense of some inference issues existing in the literature of short-term health effects of air pollution. Yet, this analysis does not allow us to clearly identify the design parameters

causing these issues. Design parameters, such as the number of observations for instance, were fixed for each study. We only observed cross-study variations and could not observe how power would evolve with the value of a parameter, for a given study. We therefore undertake a prospective design analysis to overcome this limitation (Altoè et al. 2020). This analysis aims to help us understand how power, type M and type S errors are affected by the value of different parameters and if some parameters play a more critical role than others.

Our prospective analysis is based on real-data simulations, only adding a treatment effect into the data. We emulate the main identification methods used in the literature and vary the treatment effect size, the number of observations, the proportion of treated units and the distribution of the outcome. We then try to analyze what could be improved in current practices by replicating exact study designs found in the literature. In all our simulations, the average of the estimates is equal to the true effect we set: we do not add any type of bias. In the present section, we describe how we implement these simulations. We first present the quasi-experiments and identification methods considered before discussing the overall setting for the simulations. We then briefly describe the data used.

4.1 Quasi-experiments and identification methods

Several methods have been implemented to estimate the short term health effects of air pollution. Researchers typically either estimate a dose response or run reduced-form analyses taking advantage of random shocks in air pollution levels or exposure.

In the former case, researchers estimate variations in the health outcome of interest as a function of air pollutant concentration. They estimate linear models, regressing the health outcome of interest on the level of pollution, controlling for variables such as weather parameters, calendar and city fixed effects. In the present analysis, we gather such analyses under the umbrella term "OLS" as they are often estimated using Ordinary Least Squares (OLS).

Another part of the literature, to avoid potential endogeneity issues, instrument the level of pollution using thermal inversions (Beard et al. 2012, Arceo et al. 2016, Jans et al. 2018), wind direction or speed (Schwartz et al. 2018, Deryugina et al. 2019, Isphording and Pestel 2021), variations in transport traffic (Moretti and Neidell 2011, Knittel et al. 2016, Schlenker and Walker 2016) or extreme natural events such as sandstorms or volcano eruptions (Ebenstein et al. 2015, Halliday et al. 2019).

In our simulations, we simplify the analysis and only consider binary instruments such as the presence of a thermal inversion or not, high/low wind speed, presence of traffic congestion or not. This assumption is not only helpful but also realistic as several papers exploit binary instruments. We randomly allocate the treatment and artificially increase the level of pollution accordingly. We then estimate the effect of air pollution on the health outcome of interest using Two Stage Least Squares (2SLS). The key assumption is that the instrument only affects the health outcome via its effect on air pollution. This assumption is verified in our simulations.

The reduced-form literature mainly studies the sparse shocks such as public transportation strikes or air pollution alerts. Researchers estimate the effect of the treatment without modeling its impact on air pollution level. We model these shocks as random events, occurring with a given probability on each day. We estimate the effect of the treatment using a simple linear model with fixed effects. This yields the Average Treatement Effect (ATE).The main identification assumption is the independence assumption, *i.e.*, that potential outcomes are independent of the treatment. In our simulations, this assumption holds since the treatment is allocated randomly.

In some cities, when air pollution levels reach a given threshold, air pollution alerts are released. This treatment can reduce both exposure to air pollution and levels of pollution. We estimate the effect of this type of intervention without modeling its impact on air pollution. To do so, we consider a Regression Discontinuity Design (RDD). The overall idea of the RDD is to compare days just below the threshold to days just above it. Just above the threshold, exposure and health impacts should be lower due to avoidance behavior or decreased pollution in response to the alert. The key identification assumption is that days just below and just above the threshold are comparable. Thus, no confounders should vary discontinuously at the threshold (local independence) and the treatment should vary at threshold (relevance). We model this so that both these assumptions are verified. However, for large bandwidths, observations above and below the threshold may be less comparable. This identification method enables to estimate the Average Treatment Effect at the cutoff.

4.2 Simulations Set-Up

Our simulations are implemented as follows:

1. Draw a study period,

2. Draw treated days, if any,
3. Create the health outcome based on the treatment effect,
4. Run the estimation,
5. Store the point estimate of interest and its standard error Repeat the procedure 1000 times,
6. Compute the power, type M, type S errors.

To be more specific, the study period is drawn at random. A given number of cities and days are drawn from the data set. We consider the same study period for each city. This seems realistic as studies focusing on several cities typically consider a unique study period. The drawing procedure for treated days depends on the quasi-experiment considered and the proportion of treated observations desired. For a treatment on random days, the treatment status for each day is drawn from a Bernoulli distribution with parameter equal to the proportion of treated observations desired. 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. The generative process for the health outcome depends on the identification method. For the reduced form approach (and RDD), the treatment effect is drawn from a Poisson distribution with parameter corresponding to the desired effect size. For the OLS, we build a generative model that creates fake health data based on the model considered and with an effect corresponding to the desired effect size. For the IV, we use the same method as for the OLS but modify the value of pollutant concentration through the instrument: $Poll_{ct}^{fake} = Poll_{ct} + \delta T_{ct} + e_{ct}$, where T_{ct} is the treatment dummy for city c at time t, δ the instrument strength and $e \sim \mathcal{N}(0, 0.1)$ noise. The 0.1 standard deviation for the noise is arbitrary but chosen so that the resulting noise is not too large.

4.3 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. 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 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 recorded in the most cities over the period and is correlated with the concentration of other pollutants. Our simulation analysis is not affected by this choice of air pollutant. 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).

5 Results

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

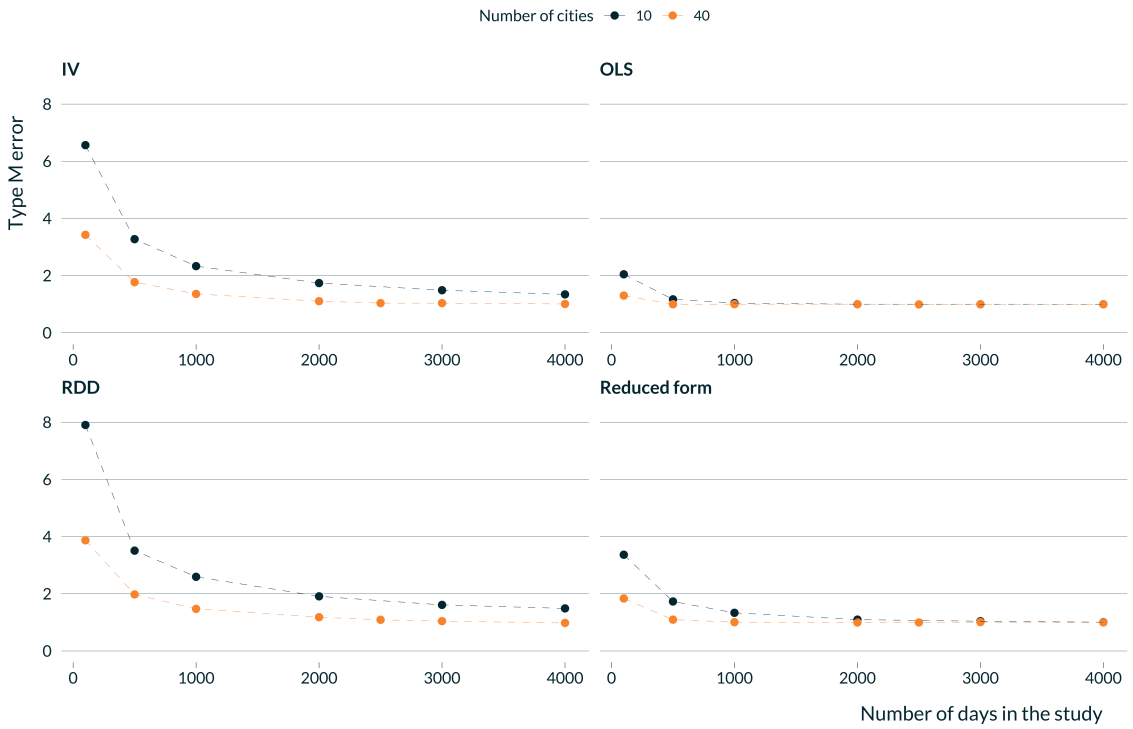
First, we analyze how power, type M and S errors are affected by the value of different design parameters. To do so, we set baseline values for these parameters and vary the value of each of them one by one. This enables us to get a sense of the impact of each parameter *ceteris paribus*. We choose relatively advantageous baseline values. We pick a large number of observations: 100,000 observations (2500 days and 40 cities). We consider a large baseline effect, as compared to the standard literature: a 1 standard deviation increase in air pollutant concentration or the occurrence of treatment leads to a 1% increase in the health outcome. For the treatment on random days, we consider an optimal proportion of treated units: 50%. For air pollution alerts, we choose a smaller but realistic proportion of treated units: 10%. We also use the largest health outcome possible in the baseline: the total number of deaths. We consider a model with as many control variables as possible: temperature, temperature squared, city fixed effects and month, year, month×year, weekday fixed effects. We also repeat this analysis for a smaller number of observations, more representative of the literature: 10,000 observations (1000 days and 10 cities).

5.1.1 Sample Size

For all identification methods, power increases and type M error decreases with the number of observations. This can be explained by an increase in the precision of the estimates. Figure 7 illustrates part of this relationship.

Importantly, even though all parameters are set to be rather advantageous, power and type M issues arise even for a large number of observations. For 40 cities and 1000 days, statistically significant estimates overestimate the effect by a factor 1.36

Figure 7: Evolution of type M Error with the number of days, comparison across identification methods



Notes: Effect size: 1%, outcome: total number of deaths, proportion of treated observations: 0.5 in the case of the IV and reduced form and 0.1 for the RDD.

(the type M error) in the case of the IV and 1.47 for the RDD. Power is respectively 53.9% and 56.1%. We cannot directly compare the RDD with the other identification methods as, for realism concerns, we set the proportion of treated units to be much smaller. The OLS seems to be much less prone to power issues than the IV. This is explained by the fact that the variance of the IV estimator is larger than the variance of the OLS estimator. For 40 cities and 1000 days, power is equal almost equal to 100% for the OLS. The reduced form approach does not suffer from substantial power issues. However, here we set parameters, in particular the proportion of treated units, to correspond to the ideal case of an RCT. In actual studies, the proportion of treated units is much smaller.

We also note that, for all identification method, Type S error is not a problem for any sample sizes.

A side analysis also shows that the distribution between the number of days and the number of cities does not matter for power—only the total number of observa-

tions does. Changing the ratio between the number of cities and days while holding the number observations constant does not affect power, type M nor type S error.

5.1.2 Effect Size

For all identification methods, the larger the effect size, the larger the power and the lower type M and S errors are. Larger effect sizes, *ceteris paribus* are associated with larger signal to noise ratio and therefore lower type M and S errors. Even with advantageous parameters, power issues start to appear for effect sizes below 1%, both for the IV and the RDD. For instance, for an effect of 0.5%, type M errors for these identification methods are about 1.7. Such effect sizes are not far off the ones found in the standard literature. With the smaller but still reasonably large dataset, power issues arise even for larger effect sizes with type M error respectively equal to 1.33 and 1.45 for a 2% effect size. With the parameters chosen, the OLS and RCT-like identification strategies seem to suffer less from power issues, even for small effect sizes. Type M error starts to increase and power to fall only for very small effect sizes. Type S error does not seem to be a key issue for any of the identification methods, even for very small effect sizes.

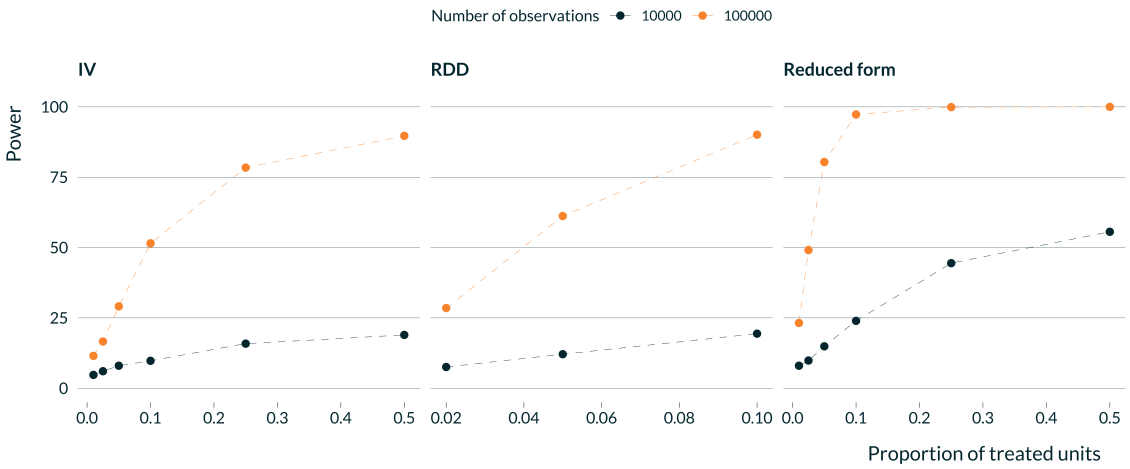
5.1.3 Proportion of Treated Units

Power decreases sharply with the proportion of treated unit for all identification methods, as visible in figure Figure 8.

The link between proportion of treated units and power might be slightly less intuitive than for sample size or effect size. Treatment effect is identified on units where treatment status changes. The size of this group decreases when the proportion of treated units decreases, leading to less a precise estimation of the treatment effect and therefore a lower power. As a consequence, type M error increases when the proportion of treated units decreases, as visible in Figure 9.

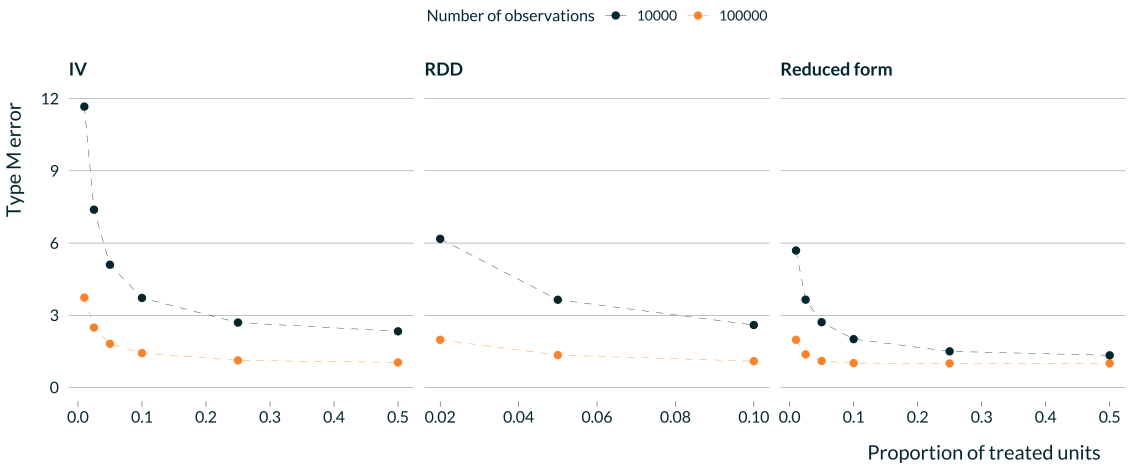
In the literature, the proportion of treated units in reduced form analyses, *i.e.*, air pollution alerts or transportation strikes, are very small, often less than 5%. Even with a large data set and rather advantageous parameters values, when the proportion of treated units is equal to 5%, type M error reaches 1.34 in the case of the RDD and 1.1 for the reduced-form analyses. With the smaller data set, it reaches 3.6 and 2.7 respectively. This literature might therefore be particularly prone to type M error due to a very low proportion of treated units, even though sample sizes are large. IV analyses might also suffer from such an issue as, with the large

Figure 8: Evolution of power with the proportion of treated units, comparison across identification methods



Notes: Effect size: 1%, outcome: total number of deaths.

Figure 9: Evolution of type M error with the proportion of treated units, comparison across identification methods



Notes: Effect size: 1%, outcome: total number of deaths.

data set and 10% of treated units, type M error is equal to 1.4.

5.1.4 Average Count of Cases of the Health Outcome

We analyze whether power issues depend on the average count of cases of the outcome. For instance, a 1% increase in the number of deaths may be more difficult

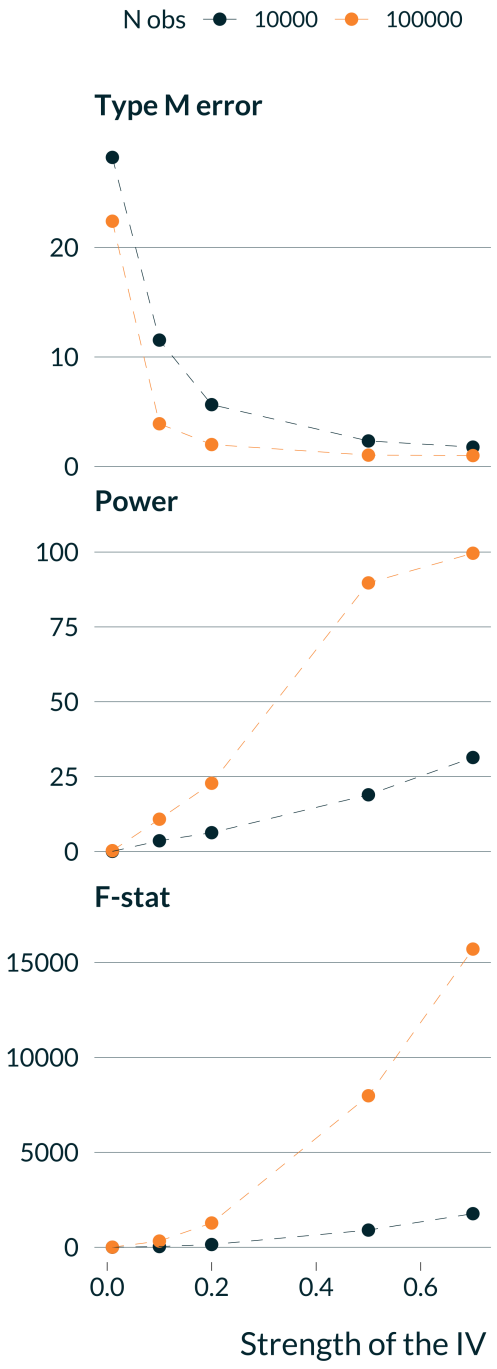
to detect when the average number of deaths is low. A 1% increase in the number of deaths in a setting where there are only 2 deaths per day corresponds to rare additional deaths that might therefore be more difficult to identify. To emulate situations with various number of cases, we consider three different outcome variables, with different counts of cases: the total number deaths, from of all causes excluding accidents (mean $\simeq$ 23 deaths per day and per city), the total number of respiratory deaths (mean $\simeq$ 2) and the number of chronic obstructive pulmonary disease cases for people aged between 65 and 75 (mean $\simeq$ 0.3).

Less intuitively than sample and effect sizes, the average count of cases critically affects power. In the large data set, while for baseline parameters and considering the total number of deaths, power is close to 100% for all identification method, when considering the total number of respiratory diseases, power falls to 11.2%, 15.8% and 50.4% for the RDD, the IV and the reduced form respectively. The corresponding type M errors are 2.80, 2.44 and 1.37 respectively. It increases to 6.19, 5.88 and 2.92 when considering the last outcome variable. Situations with a small count of cases may therefore lead to extreme type M error and power issues.

5.1.5 Issues Specific to the Instrumental Variable Design

For the instrumental variable identification strategy, we also analyze how power is affected by the strength of the instrument. The strength of the instrument is defined as the magnitude of effect of the instrument on CO concentration (refer to section 4.2). We find that, power collapses and type M error soars when IV strength decreases, as visible in Figure 10. Importantly, this issue arises for rather large IV strengths. Even in the case of the large data set, for an IV strength of 0.2, power is only 22.8% and type M equal to 2.0. This issue arises even when F-stats are large. In the case described above, the F-stat is huge and equal to 1278. This is way past the usual threshold below which we usually consider that the IV is weak, 10. This large F-stat despite the limited strength of the IV may be explained by the fact that the F-stat directly depends on the sample size.

Figure 10: Evolution of power, type M and F-stat with the IV strength



Notes: Effect size: 1%, outcome: total number of deaths, proportion of treated units: 50%

5.2 Simulating Current Practices in the Causal Inference Literature

Our simulations enable us to study how power, type M and type S evolved with the value of various parameters; They represent an "ideal" setting, with relatively large sample and effect sizes, proportion of treated units, outcome count

and IV strength. These parameters may not perfectly represent actual studies. For each identification method, we therefore consider a given 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 falling into power issues.

5.2.1 Transportation strikes

Transportation strikes are rare events. Therefore, 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) study five cities over a period of 6 years, for a total number of 11,000 observations but only observe 45 1-day strikes over this period. The proportion of treated units is therefore of 0.4%. Based on the results described in the previous section, this proportion is concernedly low and might be associated with high risks of type M error. We therefore simulate a similar design in order to get a sense of potential type M error. In our baseline simulation, we consider that the true effect is equal to the one they find (11%). We also consider an health outcome with a average count of cases close to theirs. We actually take a conservative approach and consider one that is 3 times larger than what is observed in their study (the total number of respiratory deaths, with a mean of 1.98 as compared to 0.692 in their study).

In the baseline scenario, power is only 14.5%. Importantly, this value is obtained for a very large true effect size. Even with such a true effect size, we find that, on average, a statistically significant estimate overestimates the true effect by 2.73 (type M error). Bauernschuster et al. might therefore greatly overestimate the true effect size. If the true effect size is smaller, type M error would be even larger.

The key limiting factor in this analysis is the extremely small proportion of treated units. If it was larger, type M would be lower but still problematic [give numbers]. If the analysis would have been carried out in larger cities with a larger count of deaths, for instance about 23 deaths per day, everything else equal, the type M would not be a problem, assuming that the true effect size would be 11%. However, we have seen that the true effect size is likely smaller than 11%. If the true effect was only 5%, type M error would be equal to 1.84. Even with a larger count of deaths Bauernschuster et al. would certainly fall into power issues.

Even though we focused on a particular study, other papers in this literature display similar characteristics, in particular a small proportion of treated units. This is inherent to the treatment as transportation strikes are rare events in most setting.

This may lead to an overestimation of the true effect.

5.2.2 Air pollution Alerts

Air pollution alerts are also rare events. In addition, their effect is generally estimated using RDD, an identification strategy that only consider observation close to the threshold. As a consequence, effective data sets may end up being particularly small in this literature as well. For instance, in Chen et al. (2018), while the initial sample size is of 3652 observations, the effective sample size is only of 143 (100 control observations and 43 treated ones). The proportion of treated observation is therefore particularly small in this case (1.2%). As for transportation strikes, we replicate the setting of Chen et al. (2018). In the baseline, we consider one cities, 3652 days, a proportion of treated of 1.2%. We also consider a true effect size of 12%, as found in the study. The average number of case in the study is 26 cases per day. We therefore use the total number of deaths as our outcome variable (mean of 23.36 deaths per day and per city).

In the baseline scenario, power is only 10.2% and type M error 4.6. This is extremely preoccupying. The estimate effect is likely to greatly overestimate the true effect, even if the true effect was as large as 12%. If we consider smaller true effect sizes, type M error shoots up and power collapses. As a consequence, we cannot really have any confidence in the reported effect sizes as statistically significant estimates overestimate true effect sizes.

5.2.3 Instrumenting Air Pollution

Papers published instrumenting air pollution often present very large data sets. For instance Schwartz et al. (2018) consider 591,570 observations (135 cities with a length of study of approximately 4382 days). In this case, air pollution is instrumented with a complex mix of variables and we cannot easily observe the proportion of treated units. However, as seen in section 5.1.5, the strength of the IV may play a crucial role in the potential presence of power issues. We therefore simulate their analysis, varying the strength of the IV. The effect size is 1.5% and the average case count 22.8. We thus use the total number of 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.

Considering a conservative IV strength of 0.5, we find a power of nearly 100% and a type M error of 1. Yet, for smaller values of the IV strength parameter, type

M quickly rises and power decreases. For an IV strength of 0.2 or 0.1, power falls to 48.3% and 16.4% respectively while type M reaches 1.43 and 2.61. Yet, in these cases the F-stat remains extremely large, 1287 and 320 respectively, and may hide these 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. Until now, most researchers ignore inference issues: few power formulas are available and the risk of type M error is largely unknown. Our retrospective analysis of the literature proves that under-powered studies with inflated effect sizes are an actual issue. In the causal inference literature, several papers appear to be dramatically under-powered and are likely to overestimate their effects by a factor of at least 1.5! One of these problematic papers even has a sample size of 73 million observations. In the standard literature, a non negligible share of studies have enough statistical power to overcome the type M error issue. Yet, a substantial share of papers does not, revealing a broad lack of concern for these issues. Noticeably, the fraction of under-powered studies has remained constant over time. We thus urge researchers to add retrospective calculations to their toolbox. They are very easy to implement and force researchers to reflect on the range of plausible effect sizes they are trying to estimate.

Unfortunately, a retrospective analysis does not help researchers understand which parameters of the research design influence the power of their studies. Our prospective analysis, using simulations based on real-data, fills this gap. Despite their large sample sizes, researchers exploiting rare exogenous shocks such as transport strikes should be aware that the small proportion of "treated" units observed in their studies can lead to a dramatically low statistical power. Instrumented dose-responses can also be problematic if the correlation between the instrument and the air pollution is limited. Two-stage least square estimates are more prone to type

M error than the "naive" ordinary least square estimate. This could question the benefit of using an instrumental variable strategy if one thinks that the amount of omitted variable is small. The regression-discontinuity design applied to air pollution alerts particularly stands out in terms of inference issues. Given the sample size its entails, we advice researchers to interpret findings with extra-care as type M error can be extremely large. Last but not least, the power of all research designs is influenced by the average count of the health outcome. For instance, in the causal inference literature, many articles investigate the acute effects of air pollution for sub-populations such as children. In such settings, there a huge risk to make a type M error, even with large sample sizes.

On top of these specific guidelines, we insist on three general recommendations to reform current research practices. First, as it has been advocated as early as the 2000s in economics by Ziliak and McCloskey (2008) and more recently by Gelman et al., researchers should abandon the null hypothesis testing framework. In the causal inference literature, 77% of the articles dichotomize evidence using the 5% threshold. Due to current editorial policies, "statistically insignificant" results are likely to be kept in the file drawer whereas published "statistically significant" estimates could be inflated (Ioannidis 2005). Second, researchers should display, along with their results, 95% and 99% confidence intervals. These intervals give the set of effects sizes supported by the data. The interpretation of the lower and upper bounds of the confidence intervals should force researchers to evaluate the precision of their estimate. Our third and last recommendation is also the most difficult to implement: studies should be replicated with identical research designs and in similar contexts. Replication exercises are still under-valued academically but would be instrumental to get a sense of the actual distribution of acute health effects of air pollution.

We hope that our article reminds us that a credible identification strategy does not necessarily yield a correct estimation of the actual true effect. 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 policy-makers evaluate the adverse effects of air pollution.

References

Gianmarco Altoè, Giulia Bertoldo, Claudio Zandonella Callegher, Enrico Toffalini, Antonio Calcagnì, Livio Finos, and Massimiliano Pastore. 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.

Eva Arceo, Rema Hanna, and Paulina Oliva. 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.

Stefan Bauernschuster, Timo Hener, and Helmut Rainer. 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.

John D. Beard, Celeste Beck, Randall Graham, Steven C. Packham, Monica Traphagan, Rebecca T. Giles, and John G. Morgan. 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.

Marie-Abèle Bind. 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.

Bernard Black, Alex Hollingsworth, Leticia Nunes, and Kosali Simon. The Effect of Health Insurance on Mortality: Power Analysis and What We Can Learn from the Affordable Care Act Coverage Expansions. Technical Report w25568, National Bureau of Economic Research, Cambridge, MA, February 2019.

Hong Chen, Qiongsi Li, Jay S Kaufman, Jun Wang, Ray Copes, Yushan Su, and Tarik Benmarhnia. 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.

Tatyana Deryugina, Garth Heutel, Nolan H. Miller, David Molitor, and Julian Reif. 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.

Francesca Dominici and Corwin Zigler. 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.

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

Andrew Gelman and John Carlin. 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.

Andrew Gelman and Francis Tuerlinckx. 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.

Andrew Gelman, Jennifer Hill, and Aki Vehtari. *Regression and Other Stories*.

Timothy J Halliday, John Lynham, and Áureo de Paula. 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.

John P. A. Ioannidis. Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8):e124, August 2005. ISSN 1549-1676. doi: 10.1371/journal.pmed.0020124.

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

Ingo E. Isphording and Nico Pestel. 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.

Jenny Jans, Per Johansson, and J. Peter Nilsson. 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.

Christopher R. Knittel, Douglas L. Miller, and Nicholas J. Sanders. 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.

Cong Liu, Renjie Chen, Francesco Sera, Ana M. Vicedo-Cabrera, Yuming Guo, Shilu Tong, Micheline S.Z.S. Coelho, Paulo H.N. Saldiva, Eric Lavigne, Patricia Matus, Nicolas Valdes Ortega, Samuel Osorio Garcia, Mathilde Pascal, Massimo Stafoggia, Matteo Scortichini, Masahiro Hashizume, Yasushi Honda, Magali Hurtado-

Díaz, Julio Cruz, Baltazar Nunes, João P. Teixeira, Ho Kim, Aurelio Tobias, Carmen Íñiguez, Bertil Forsberg, Christofer Åström, Martina S. Ragettli, Yue-Leon Guo, Bing-Yu Chen, Michelle L. Bell, Caradee Y. Wright, Noah Scovronick, Rebecca M. Garland, Ai Milojevic, Jan Kyselý, Aleš Urban, Hans Orru, Ene Indermitte, Jouni J.K. Jaakkola, Niilo R.I. Ryti, Klea Katsouyanni, Antonis Analitis, Antonella Zanobetti, Joel Schwartz, Jianmin Chen, Tangchun Wu, Aaron Cohen, Antonio Gasparrini, and Haidong Kan. 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.

Jiannan Lu, Yixuan Qiu, and Alex Deng. A note on Type S/M errors in hypothesis testing. *British Journal of Mathematical and Statistical Psychology*, 72(1):1–17, 2019. ISSN 2044-8317. doi: 10.1111/bmsp.12132.

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

Enrico Moretti and Matthew Neidell. 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.

Wolfram Schlenker and W. Reed Walker. 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.

Joel Schwartz, Elena Austin, Marie-Abele Bind, Antonella Zanobetti, and Petros Koutrakis. 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.

Joel Schwartz, Kelvin Fong, and Antonella Zanobetti. 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.

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

Stef van Buuren and Michael Greenacre. Flexible Imputation of Missing Data, Second Edition. page 444.

Nuntavarn Vichit-Vadakan, Nitaya Vajanapoom, and Bart Ostro. 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.

Stephen Thomas Ziliak and Deirdre N. McCloskey. *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, 2008. ISBN 978-0-472-07007-7 978-0-472-05007-9.