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

Vincent Bagilet[1]

Léo Zabrocki-Hallak[2]

November 16th, 2022

Abstract

Hundreds of studies have shown that air pollution affects health in the very short-run. This played a key role in setting air quality standards. Yet, estimated effect sizes can vary widely across studies. Analyzing the results published in epidemiology and economics, we find that publication bias and a lack of statistical power could lead some estimates to be inflated. We then run real data simulations to identify the design parameters causing these issues. We show that this exaggeration may be driven by a small numbers of exogenous shocks, instruments with limited strength or sparse outcomes. Other literatures relying on comparable research design could also be affected by these issues. Our paper provides a principled workflow to evaluate and avoid the risk of exaggeration when conducting an observational study.

Website: https://vincentbagilet.github.io/inference_pollution

Replication Materials: https://osf.io/p6725/

[†]**Comments and suggestions are highly welcome.** We are very grateful to Hélène Ollivier and Jeffrey Shrader for their guidance on this project. Many thanks to Michela Baccini, Geoffrey Barrows, Tarik Benmarhnia, Marie-Abèle Bind, Sylvain Chabé-Ferret, Clément de Chaisemartin, Tatyana Deryugina, Ludovica Gazze, Marion Leroutier, Quentin Lippmann, Jesse McDevitt-Irwin, Claire Palandri, Thomas Piketty, as well as seminars participants at Columbia SusDev Colloquium, PSE, IPWSD, Benmarhnia's Lab, M&A's Lab, FAERE, and EuHEA for their feedbacks.

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

[2]RFF-CMCC European Institute on Economics and the Environment, Milan, Italy. Email: leo.zabrocki@gmail.com

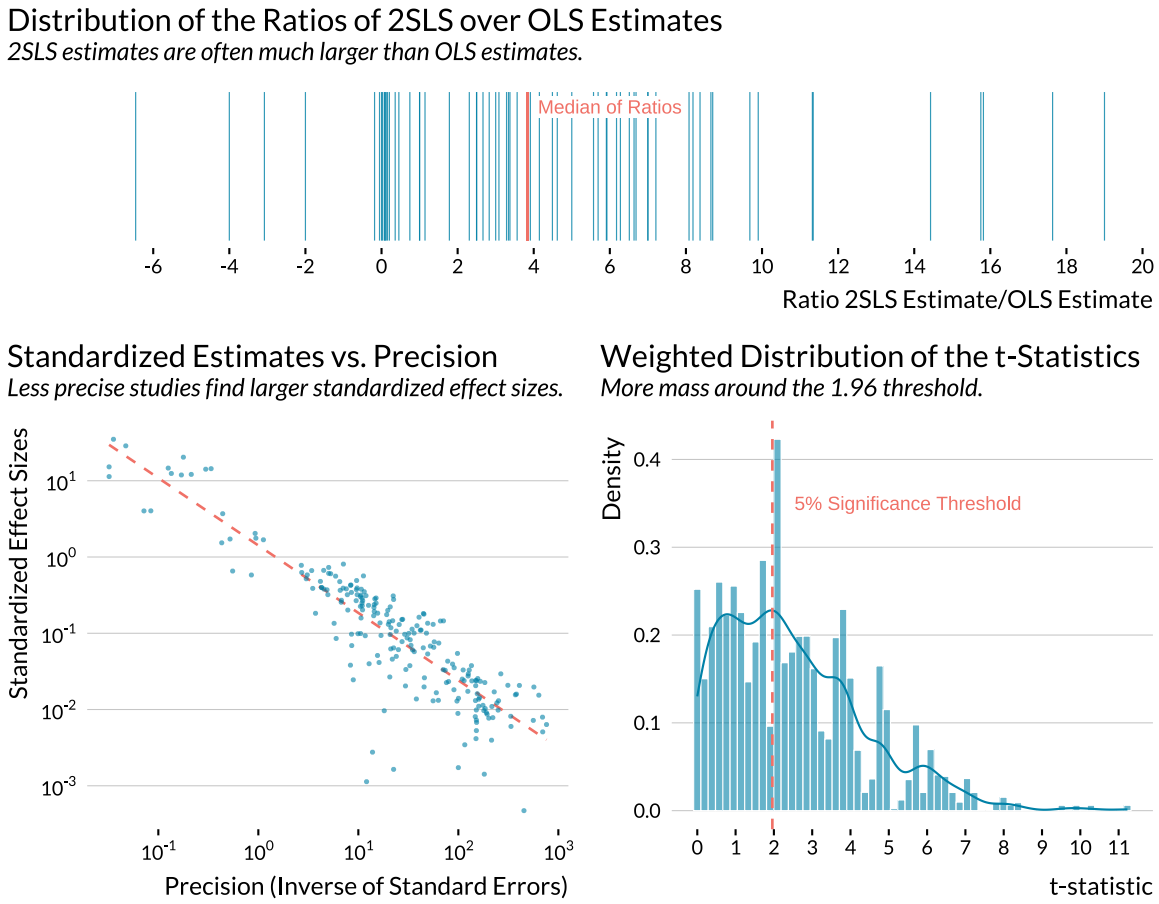
1 Introduction

From extreme events such as the London Fog of 1952 to the development of sophisticated time-series analyses, a vast epidemiology literature of more than 600 studies has established that air pollution induces adverse health effects on the very short-term. Increases in the concentration of several ambient air pollutants have been found to be associated with small increases in daily mortality and emergency admissions for respiratory and cardiovascular causes (Schwartz 1994, Samet et al. 2000, Le Tertre et al. 2002, Bell et al. 2004, Liu et al. 2019). Based on these results, environmental protection and public health agencies have designed policies such as air quality alerts to mitigate the burden of air pollution. Obtaining accurate estimates is therefore crucial as they are directly used to implement and update policies.

With this objective in mind, researchers in economics and epidemiology have recently used causal inference methods to improve on the standard epidemiology literature that relied on associations (Dominici and Zigler 2017, Bind 2019). Newly obtained results confirm the short-term health effects of air pollution (Schwartz et al. 2015; 2018, Deryugina et al. 2019). Yet, causal estimates are up to an order of magnitude larger than what would have been predicted by the standard epidemiology literature. Reviewing the causal inference literature, we find that the median of the ratio of Two-Stage Least-Squares (2SLS) to "naive" Ordinary Least-Squares (OLS) estimates is 3.8, as shown in the top panel of Figure 1. This discrepancy could arguably be explained by the fact that instrumental variable strategies remove omitted variable bias, reduce attenuation bias caused by classical measurement error in air pollution exposure or target a different causal estimand.

Our literature review however suggests an alternative but complementary explanation based on publication bias and low statistical power. The bottom left panel of Figure 1 reveals that large standardized effect sizes are only found in imprecise stud-

Figure 1: Suggestive Evidence of Publication Bias, Power and Exaggeration Issues in the Causal Inference Literature.



Notes: In the top panel, we plot the ratios of the 2SLS estimate to the corresponding "naive" OLS one (same health outcome and air pollutant). Out of the 72 ratios obtained, we exclude 7 outliers with extremely large ratios that distort the graph. The orange line represents the median and is equal to 3.8. In the bottom left panel, we display 382 standardized effect sizes against the inverse of their standard error, a measure of precision. Both axes are on a log10 scale. In the bottom right panel, following Brodeur et al. (2020), we plot the weighted distribution of the 537 t-statistics. The weights are equal to the inverse of the number of tests displayed in the same table multiplied by the inverse of the number of tables in the article. The dashed orange line represents the 5% significance threshold.

ies. This pattern is often indicative of publication bias. Among imprecise studies, those that found an effect size large enough to be statistically significant—at least 2 standard errors away from 0 at the 5% significance level—were more likely to be published (Brodeur et al. 2016; 2020). Studies with low precision therefore produce inflated estimates in the presence of publication bias (Ioannidis 2008, Gelman and Carlin 2014).

The bottom right panel of Figure 1 confirms the presence of a publication bias in this literature as the mass of the t-statistics distribution is larger at the 5% statistical significance threshold. While many other literatures do suffer from exaggeration issues, the consequences of low statistical power are particularly salient in studies on the short-term health effects of air pollution since their signal-to-noise ratio is often low (Peng et al. 2006, Peng and Dominici 2008).

In this paper, we analyze the consequences and determinants of low statistical power in studies on the short-term health effects of air pollution. We first tackle this question by gathering a unique corpus of 668 articles based on associations and 36 articles that rely on causal inference methods. For each of these studies, we run statistical power calculations to assess whether the design of the study would be robust enough to capture the true effect if it was smaller than the observed estimate (Gelman and Carlin 2014, Ioannidis et al. 2017, Lu et al. 2019, Timm 2019). Yet, these calculations rely on hypotheses about the true effect of the treatment and do not enable to understand the causes of low power. Using real data from the US National Morbidity, Mortality, and Air Pollution Study (Samet et al. 2000), we therefore implement simulations to identify the characteristics of research designs that drive their statistical power and the inflation of statistically significant estimates (Altoè et al. 2020, Gelman et al. 2020, Black et al. 2022). We finally provide a principled workflow to evaluate the risk of exaggeration along with a list of concrete recommendations to improve studies designs.

Our analysis of estimates published in the epidemiology and causal inference literatures show that, reassuringly, many studies are robust to low statistical power issues. These studies could recover effect sizes that are equal to 3/4 of the obtained estimate. However, a quarter of studies is likely to suffer from important exaggeration issues. Their estimated effect sizes are probably at least inflated by a factor of 1.4. Better informed guesses of true effect sizes suggest that exaggeration issues might be even more

widespread.

Our simulation results help understand why some studies face statistical power issues. We first show that, regardless of the identification strategy used, a very large number of observations is needed to reach a sufficient statistical power. Air quality alerts being rare, observations close to the air pollution threshold are scarce. Regression discontinuity designs exploiting these alerts are bound to rely on small samples and often produce inflated estimates. Second, we find that the use of public transportation strikes or thermal inversions as exogenous shocks on air pollution can be problematic. Even if these studies can have large sample sizes, the number of shocks sometimes represents less than 1% of the observations. The variation available for identification is therefore small, leading to exaggeration, even for large true effect sizes. Third, we show that the average daily count of cases of the health outcome is a key driver of statistical power. Estimated effects of air pollution on the elderly or children can be exaggerated since there are few daily hospital admissions or deaths for these groups.

By quantifying the respective influence of parameters affecting the power of studies, we fill an important gap in the literature on the acute health effects of air pollution. There was a lack of guidance on how to design an observational study to avoid low power issues, except for generalized additive models used in the standard epidemiology literature (Winquist et al. 2012). While our simulations focus on health effects, our conclusions could be extended to studies with similar designs but investigating the impacts of air pollution on different outcomes such as criminality, cognitive skills and productivity (Herrnstadt et al. 2021, Ebenstein et al. 2016, Adhvaryu et al. 2022). More broadly, we expect studies focusing on settings with small effect sizes, a limited number of exogenous shocks or of cases in the count outcome to also be subject to power and exaggeration issues.

Our paper makes three main contributions. First, it contributes to a growing litera-

ture using retrospective power calculations to assess power and exaggeration issues in various fields (Ioannidis 2008, Gelman and Carlin 2014, Ioannidis et al. 2017, Ferraro and Shukla 2020, Stommes et al. 2021, Arel-Bundock et al. 2022). These meta-analyses help understand the recent replication crises in medicine, psychology and experimental economics (Button et al. 2013, Open Science Collaboration 2015, Camerer et al. 2018). Our analysis complements the literature by showing the existence of such issues for a major branch of health and environmental economics. The algorithm we developed to automatically review the epidemiology literature is readily available to evaluate power issues in other fields reporting point estimates and confidence intervals in plain text.

Second, existing meta-analyses do not usually discuss the determinants of the lack of power they describe. By simulating all research designs used in the literature on the short-term health effects of air pollution, we are able to overcome this key limitation. Our analysis complements three recent articles using simulations to evaluate power issues in analyses of state-level public policies on mortality outcomes (Schell et al. 2018, Griffin et al. 2021, Black et al. 2022). These studies focus on event-study designs and treatment effects happening on medium to long time scales. 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.

Third, our study provides a reproducible workflow to evaluate and address power issues when running an observational study. Compared to psychology (Altoè et al. 2020), researchers in economics lack concrete recommendations to evaluate and understand the causes of low power issues. Before carrying out the study, we suggest to build simulations using existing datasets to evaluate how the performance of the research design evolves with key parameters. Once the analysis is completed, we recommend to run and report a retrospective power analysis to assess whether the design

used would have recovered the true effect if it was in fact smaller than the one estimated. To ease the adoption of these tools, all replication and supplementary materials are available on the project's website.

In the following section, we implement a simple simulation exercise to show why statistically significant estimates exaggerate true effect sizes when studies have 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, Exaggeration and Type S Error

In a seminal paper, Gelman and Carlin (2014) point out that statistically significant estimates suffer from a winner's curse in under-powered studies. These estimates can largely overestimate true effect sizes or can even be of the opposite sign. In this section, we implement a simple simulation exercise to illustrate these two seemingly counter-intuitive issues.

2.1 Illustrative Example

We simulate an experiment in which a mad scientist is able to increase the concentration of fine particulate matter ($PM_{2.5}$) to estimate the short-term effects of air pollution on daily non-accidental mortality. The experiment takes place in a major city over the 366 days of a leap year. The scientist increases the concentration of particulate matter by 10 $\mu g/m^3$—a large shock equivalent to a 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. They find a treatment effect of 4 additional deaths that is statistically significant at the 5% level. The statistical significance of the estimate fulfills the scientist expectations.

Table 1: Science Table of the Experiment.

Day Index	$Y_i(0)$	$Y_i(1)$	τ_i	T_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 daily number of non-accidental deaths for 6 of the 366 daily units in the scientist's experiment.

Contrary to the scientist, we know the true effect of the experiment since we created the data. In Table 1, we display the pair of potential outcomes of each day, $Y_i(T_i = 0)$ and $Y_i(T_i = 1)$. $Y_i(T_i)$ represents the daily count of non-accidental deaths and T_i the treatment assignment, equal to 1 for treated units and 0 otherwise. We first simulate the daily non-accidental mortality counts in the absence of treatment (i.e., the $Y(0)$ column of Table 1), by drawing 366 observations from a negative binomial distribution with a mean of 106 and a variance of 402. We chose these parameters to approximate the distribution of non-accidental mortality counts in a large European city. We then define the counterfactual distribution of mortality by adding, on average, 1 extra death (i.e., the $Y(1)$ column of Table 1). We then define the counterfactual distribution of mortality by adding the treatment effect, drawn from a Poisson distribution (i.e., the $Y(1)$ column of Table 1). We choose its parameter to increase the number of death by

1 on average[1].

Following the fundamental problem of causal inference, the daily count of deaths the scientist observes is given by the equation: $Y_i^{\text{obs}} = T_i \times Y_i(1) + (1 - T_i) \times Y_i(0)$. Considering that the assignment of the treatment was random, how can the statistically significant estimate found by the scientist be 4 times larger than the true treatment effect size? Replicating the experiment a large number of times explains this apparently puzzling result.

2.2 Defining Statistical Power, Exaggeration ratio and Type S error

Figure 2: Replicating 10,000 Times the Experiment.

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.

[1] In relative terms, the treatment effect size we set represents a 1% increase in the health outcome. The magnitude of this hypothetical effect is larger than the one found in a recent and large-scale study based on 625 cities. Liu et al. (2019) estimated 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.

In Figure 2, we plot the estimates of 10,000 iterations of the experiment. Even if there is a large variation in the effect size of estimates, their 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, several replications of this experiment would recover the true treatment effect. Unfortunately, despite recent changes in scientific practices and editorial policies, non-statistically significant estimates and replication exercises remain not valued enough (Brodeur et al. 2020). In a world with publication bias, statistically significant estimates are more likely to be made public. Out of the 10,000 simulation estimates, about 800 are statistically significant at the 5% level. The *statistical power* of the experiment, which is the probability to reject the null hypothesis when there is actually an effect, is equal to 8%. The scientist was therefore lucky to get a statistically significant estimate.

With such a low statistical power, statistically significant estimates are however not informative for the treatment of interest. Two metrics, the average type M (magnitude) error and the probability to make a type S (sign) error help assess the negative consequences of a lack of statistical power. The exaggeration ratio, or expected Type M error, is defined as the 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 could expect their statistically significant estimates to be inflated on average by a factor of 5! We also notice in Figure 2 that a non-negligible fraction of statistically significant estimates are of the wrong sign: this proportion is the probability of making a type S error (Gelman and Carlin 2014). In this experiment, a statistically significant estimate has a 8% probability of being of the wrong sign!

Formally, the statistical power of a test is the probability of rejecting the null hypothesis $H_0 : \beta = 0$, where β is the true effect of the estimand of interest. For $\hat{\beta}$, a normally distributed unbiased estimate of β with a standard error σ, the power of the

null hypothesis test at the 5% is equal to $\Phi\left(-1.96-\frac{\beta}{\sigma}\right)+1-\Phi\left(1.96-\frac{\beta}{\sigma}\right)$, where Φ is the cumulative distribution function of the standard normal distribution. It increases with β, the true value of the effect and with the precision of the estimate, *i.e.*, when σ decreases. The exaggeration ratio is $\mathbb{E}\left(\frac{|\hat{\beta}|}{|\beta|}\,\middle|\,\beta,\sigma,|\hat{\beta}|/\sigma>1.96\right)$ and the probability to make a type S error is given by $\Pr\left(\frac{\hat{\beta}}{\beta}<0\,\middle|\,\beta,\sigma,|\hat{\beta}|/\sigma>1.96\right)$. Zwet and Cator (2021) and Lu et al. (2019) derive closed-form expressions for these quantities. They show that both the exaggeration ratio and the probability of type S error decrease with β and the precision of the estimate.

To obtain statistically significant estimates that are informative of the true value of the effect size, the scientist would need to improve the design of their study in order to increase its statistical power.

3 Retrospective Analysis of the Literature

In this section, we first describe how we run a retrospective analysis of the standard epidemiology and causal inference literatures. We then assess to what extent they could suffer from low statistical power issues.

3.1 Our Approach

The formulas for power, exaggeration ratio and type S error described in the previous section all depend on the true magnitude of the estimand of interest. The true effect is however never observed in a given study. We can overcome this limitation using a retrospective power analysis. Essentially, it addresses the following question: would the design of our study be reliable enough to retrieve the true effect if it was in fact smaller than the obtained estimate? A retrospective power analysis can be considered as a thought-experiment in which we would exactly replicate the study many times under the assumption that the true effect is different from the observed estimate.

Concretely, Gelman and Carlin (2014) propose to run simulations in which we draw many estimates from the asymptotic distribution of the estimator, a normal distribution with mean equal to the hypothesized true effect and a standard deviation equal to the standard error we obtained in our study. The statistical power is the proportion of sampled estimates that are statistically significant at the 5% level. The exaggeration ratio is computed as the average ratio of the values of statistically significant estimates over the assumed true effect size. The probability to make a type S error is the proportion of significant estimates that are of the opposite sign of the true value. In our project, we use the **R** package `retrodesign` developed by Timm (2019) that implements the closed-form analogue of these simulations (Lu et al. 2019).

To get a general overview of power issues in the standard epidemiology and causal inference literatures, we first carry out the same retrospective analysis for each study. What proportion of studies would have a design reliable enough to retrieve an effect size equal to 3/4 of the obtained estimate? On average, by what factor would statistically significant estimates be inflated? For a subset of studies, we then make more elaborate guesses about potential true values of the 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 rely 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.

To gather a corpus of relevant articles, we use the following search query on PubMed and Scopus:

```
'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"'
```

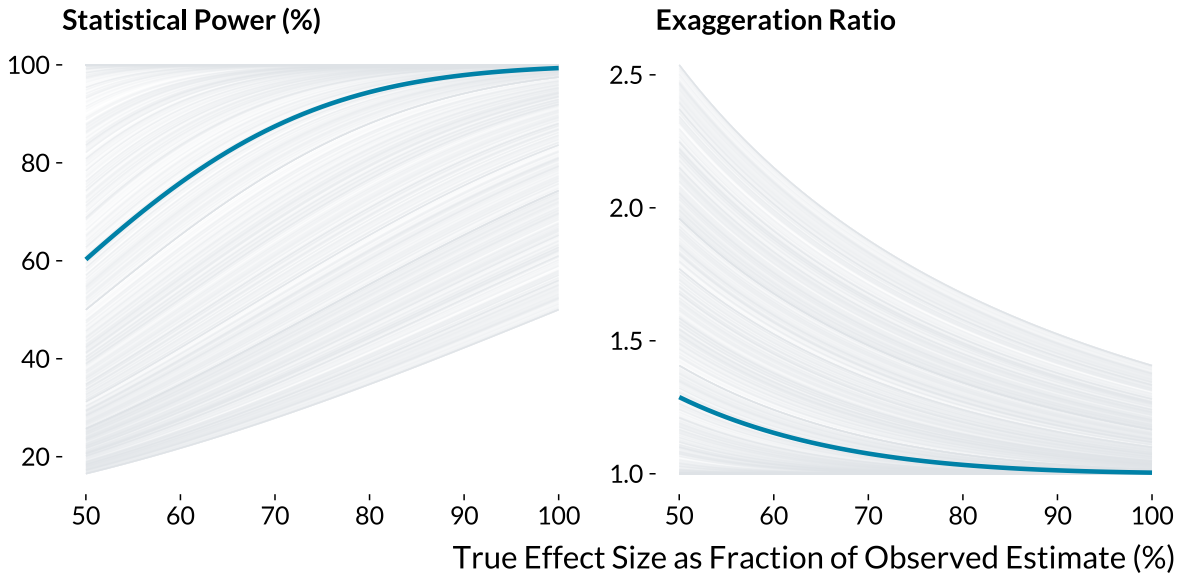
We retrieve the abstracts of 1834 articles. We then extract estimates and confidence intervals from these abstracts using regular expressions (regex). Our algorithm detects phrases such as "95% confidence interval (CI)" or "95% CI" and looks for numbers directly before this phrase or after and in a confidence interval-like format. We illustrate the outcome of this procedure (in blue) 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."

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. 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.

Based on this subset of articles, we first implement a retrospective power analysis to evaluate whether a study could recover an effect size equal to 3/4 of the obtained estimate. We carry out this analysis for the 1982 estimates that are statistically significant. In Figure 2, we display the power and exaggeration curves for each result describing how these quantities vary with the hypothetical true effect sizes. The blue lines represent the medians. If the true effect size was equal to 3/4 of the obtained

Figure 3: Power and Exaggeration Curves for the Epidemiology Literture.

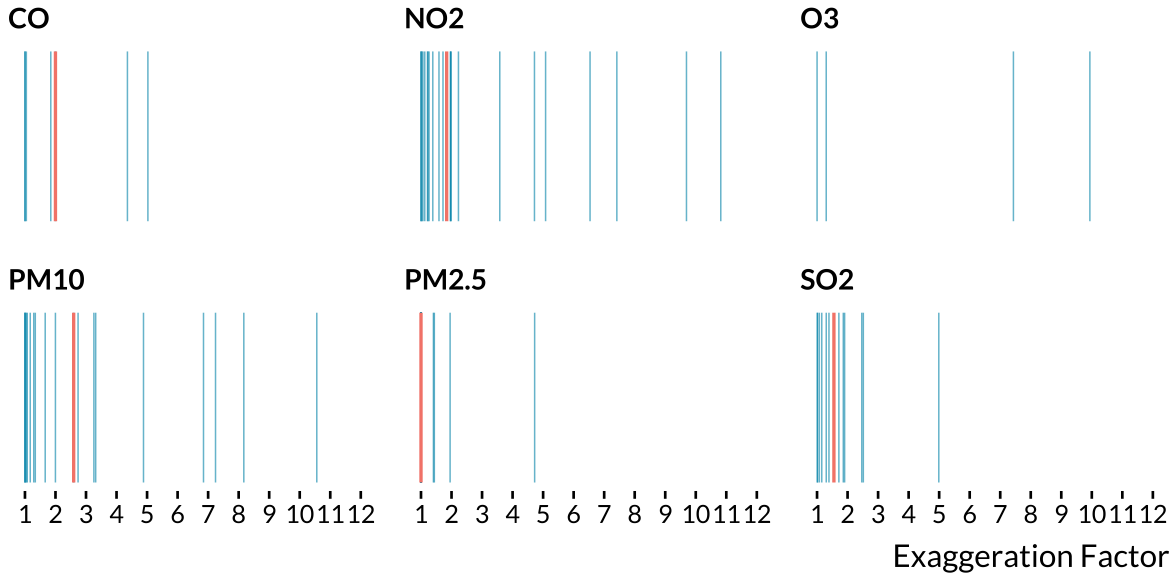


Notes: Each gray line is a power curve or an exaggeration curve of a statistically significant result published in the epidemiology literature. The blue lines are the median values. For visual clarity, we drop results for which exaggeration ratios were too large.

estimate, 57% of the studies would have a power above the conventional 80% target used in randomized controlled trials. The median exaggeration ratio would be 1.1 and type S error would not be an issue. These figures however hide a lot of heterogeneity across studies. For one quarter of studies, the exaggeration would be higher than 1.4. We therefore try to apprehend the sources of this heterogeneity.

We find that inference issues do not depend on the health outcome and the air pollutant studied. Health science journals appear to be more prone to power issues than other journals. Researchers seem to be aware that they should work with large sample size as they often carry out multi-city studies. They also sometimes explicitly state that they investigate non-accidental mortality causes to increase statistical power since the average daily count is higher than for more specific death causes. 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 exaggeration ratios.

Figure 4: Distribution of Exaggeration Ratios for Studies in Shah et al. (2015)'s Meta-Analysis.



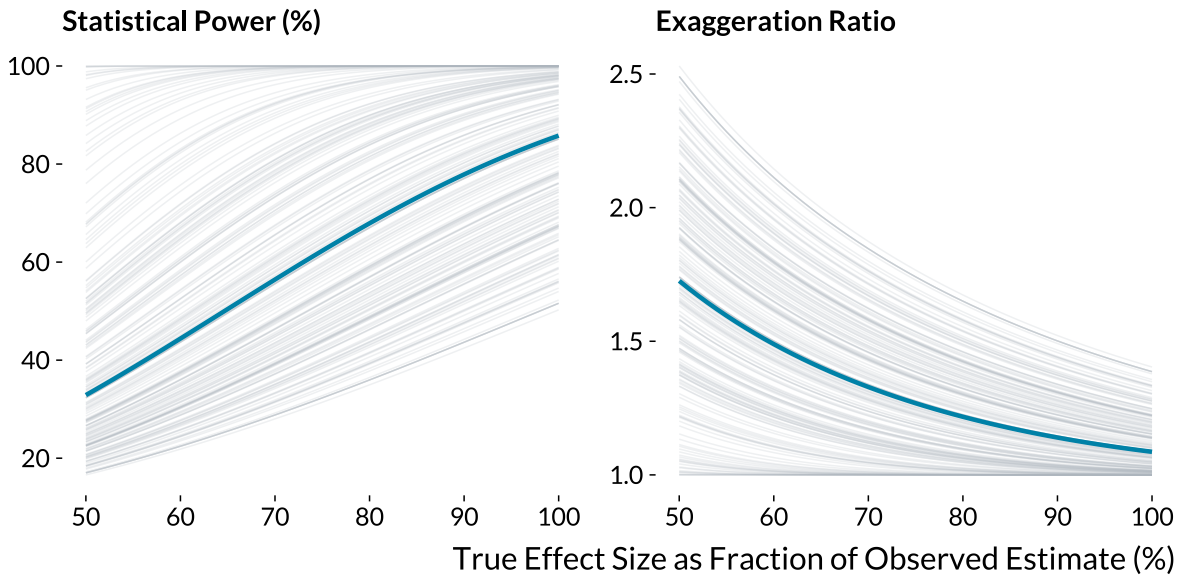
Notes: Each blue line is the exaggeration ratio of a statistically significant estimate retrieved from Shah et al. (2015)'s meta-analysis. We use the meta-analysis estimates as true effect sizes in the retrospective power calculations. Orange lines are the medians. We remove extreme exaggeration ratio for visual clarity: the median for O_3 is 13.4.

Finally, we expand our review of the standard epidemiology literature by focusing on a set of 54 studies gathered by Shah et al. (2015) on the effects of several air pollutants on mortality and emergency admission for stroke. For each of these studies, we run retrospective power calculations to evaluate their ability to retrieve the meta-analysis estimates. This approach is recommended by Gelman and Carlin (2014) and Ioannidis et al. (2017) to make more informed guesses about potential true effect sizes. 63% of studies in Shah et al. (2015) have a statistical power below 80%. The median exaggeration ratio of statistically significant estimate is equal to 1.6. In Figure 4, we plot, for each air pollutant, the distribution of the exaggeration ratios (blue lines) and their medians (orange lines). The median exaggeration varies a lot by air pollutant, from 1 for $PM_{2.5}$ up to 13.4 for O_3 (the median is not displayed for visual clarity). More informed guesses about true effect sizes confirm that exaggeration is common in the standard epidemiology literature.

3.3 Causal Inference Literature

We used an extensive search strategy on Google Scholar, IDEAS, and PubMed to retrieve studies that (i) focus on the short-health effects of air pollution on mortality and morbidity outcomes, and (ii) rely on a causal inference methods[2]. In the Appendix A.1, we display the list of the 36 articles that match our search criteria. For each study, we retrieved the method used by the authors, which health outcome and air pollutant they consider, the point estimate and the standard error. We coded the main specifications but also those on heterogeneous effects by age categories. Half of the studies report more than 11 results.

Figure 5: Statistical Power and Exaggeration Curves of Causal Inference Studies.



Notes: Each gray line is a power curve or an exaggeration curve of a statistically significant result published in the causal inference literature. The blue lines are the median values. For visual clarity, we drop results for which exaggeration ratios were too large.

To evaluate potential statistical power issues in this literature, we first proceed exactly as for the standard epidemiology literature. In Figure 5, we plot the power and exaggeration curves for 186 specifications which results are statistically significant at

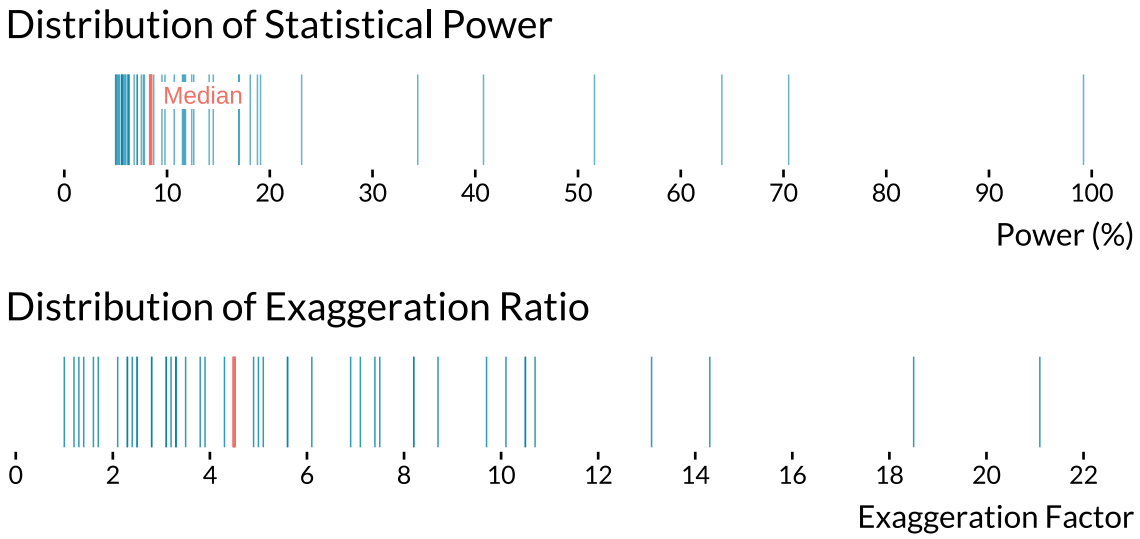
[2]We excluded the very recent literature on the effects of air pollution on COVID-19 health outcomes as we wanted to gather a relatively homogeneous corpus of studies.

the 5% level. If the true effect size of each study was equal to 3/4 of the obtained estimate, the median power would be 62% and the median exaggeration ration would be 1.3. Only 30% of studies would have a power greater than 80%. Figure 5 also shows 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 large exaggeration issues. For instance, one quarter studies would, on average, exaggerate the true effect sizes by a factor greater than 1.5. This pattern may help explain why very large effect sizes are sometimes observed in the causal inference literature.

We also supplement this general retrospective analysis with another exercise where we take as true effect sizes the estimates that would be predicted using non-causal inference methods. We focus here on instrumental variable strategies since they are the most common design in the causal inference literature. The discrepancy between OLS and 2SLS estimates is often explained by a combination of omitted variable bias and attenuation bias due to classical measurement error in air pollution exposure. It can also come from the fact that the causal estimand targeted by the naive and instrumental variable strategies are not the same if treatment effects are heterogeneous. We are however lacking evidence on the contribution of each explanation the discrepancy between non-causal and causal estimates. If we believe that omitted variable and attenuation biases are negligible, low power issues could be a part of the story.

We analyze 98 instrumental variables results for which we could retrieve the corresponding naive regression results. We take the OLS estimates as the true effect sizes in the 2SLS specifications. In Figure 6, we display the distribution of the statistical power and the average exaggeration ratio of instrumental variable results. The median power is equal to 8.4%. This results in large exaggeration ratios: half of the studies would overestimate true effect sizes by a factor of at least 4.5. Such an inflation of statistically significant estimates could partially close the gap between instrumented and non-instrumented estimates.

Figure 6: Distribution of Power and Exaggeration Ratio for Instrument Variable Designs, Assuming that the Naive OLS Estimates Are the True Effect Sizes.



Notes: For 98 statistically significant 2SLS estimates, we define the true values of effect size as the corresponding OLS estimates. Each blue line represents either the statistical power (%) or the exaggeration factor of a study's result. Orange lines are the median of the two metrics. For visual clarity, we do not display three extreme exaggeration ratios.

4 Prospective Analysis of Causal Inference Methods

The review of the standard epidemiology and causal literatures shows that some articles produce 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, Black et al. 2022). We run Monte-Carlo simulations based on real-data to emulate the main empirical strategies found in the literature. We use real data to avoid the difficult task of modeling the long-term and seasonal variations in health outcomes but also the specific effects of weather variables such as temperature. In this section, we describe how we implement

these simulations. We start by presenting the causal identification strategies used to measure the acute health effects of air pollution. We then briefly describe the data we rely on and finally detail how we implement our simulations.

4.1 Research Designs to Measure the Short-Term Health Effects of Air Pollution

Several empirical strategies have been implemented to estimate the short-term health effects of air pollution. In our simulations, we try to simulate the main ones existing in the literature. We consider a setting where data on air pollution, weather parameters, and health outcomes are aggregated at the daily-city level.

Standard regression approach. 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 they regress the daily count of an health outcome on an air pollutant concentration, while flexibly adjusting for weather parameters, seasonal and long-term variations. We approximate the workhorse model used by epidemiologists using linear models estimated via ordinary least squares:

$$\mathrm{Y}_{c,t} = \alpha + \beta \mathrm{P}_{c,t} + \mathbf{W}_{c,t}\lambda + \mathbf{C}_t\gamma + \epsilon_{c,t}$$

where c is the city index and t the daily time index. $\mathrm{Y}_{c,t}$ is the daily count of cases of an health outcome and $\mathrm{P}_{c,t}$ the average daily concentration of an air pollutant and $\epsilon_{c,t}$ an error term. The parameter β captures the short-term effect of an increase in the air pollutant concentration on the health outcome. To address confounding issues, the model adjusts for a set of weather covariates, $\mathbf{W}_{c,t}$, and calendar indicators $\mathbf{C}_t$.

Instrumental variable (IV) approach. The standard strategy could be prone to omitted variable bias and measurement error. A growing number of articles therefore exploits 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), wind patterns (Schwartz et al. 2018, Deryugina et al. 2019), 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). This approach can be summarized with a two-stage model where the first stage is:

$$\mathrm{P}_{c,t} = \delta + \theta \mathrm{Z}_{c,t} + \mathbf{W}_{c,t}\eta + \mathbf{C}_t\kappa + e_{c,t}$$

where $\mathrm{Z}_{c,t}$ is the instrumental variable. The second stage is then:

$$\mathrm{Y}_{c,t} = \alpha + \beta\widehat{\mathrm{P}}_{c,t} + \mathbf{W}_{c,t}\lambda + \mathbf{C}_t\lambda + \epsilon_{c,t}$$

where $\widehat{\mathrm{P}}_{c,t}$ is the exogenous variation in an air pollutant predicted by the instrument. The causal effect measured by this approach is a weighted average of per-unit causal responses to an increase in the concentration of an air pollutant (Angrist and Imbens 1995).

Reduced-form approach. A subset of articles directly estimates the relationship between the health outcome and exogenous shocks to air pollution. For instance, articles using this approach exploit public transport strikes or thermal inversion as exogenous shocks (Bauernschuster et al. 2017, Jans et al. 2018, Godzinski et al. 2019, Giaccherini et al. 2021):

$$\mathrm{Y}_{c,t} = \alpha + \beta\mathrm{D}_{c,t} + \mathbf{W}_{c,t}\lambda + \mathbf{C}_t\gamma + \epsilon_{c,t}$$

where $D_{c,t}$ is a dummy equal to 1 when city c is affected by a shock at time t and 0 otherwise. The parameter β captures an intention-to-treat effect.

Regression-discontinuity design (RDD) approach. The last empirical strategy found in the literature measures the effects of air quality alerts with a regression-discontinuity design (Chen et al. 2018). In this approach, the following model is estimated for observations within an air pollution concentration bandwidth around the air pollution alert threshold:

$$Y_{c,t} = \alpha + \beta \mathbf{1}\{P_{c,t} > P_c^{(a)}\} + \mathbf{W}_{c,t}\lambda + \mathbf{C}_t\gamma + \epsilon_{c,t}$$

where $P_c^{(a)}$ is the air pollution alert threshold for city c. We restrict our simulations to the case of sharp RDD. This model estimates the intention-to-treat effect of air quality alerts. It can both capture the effect of a subsequent decrease in air pollution caused by traffic restriction policies and inhabitants' avoidance behavior.

4.2 Data

Our simulation exercises are based on a subset of the US National Morbidity, Mortality, and Air Pollution Study (NMMAPS). The dataset is publicly available and has been used 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). Specifically, we extract daily data on 68 cities over the 1987-1997 period. It corresponds to 4,018 observations per city, for a total sample size of 273,224 observations. We select observations on the average temperature (C°), the standardized concentration of carbon monoxide (CO), and mortality counts for several causes. We focus on 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. We impute them using the chained random forest algorithm implemented in the

`missRanger` package (Mayer 2019).

4.3 Simulations Set-Up

General procedure. Our simulation procedure therefore follows 7 main steps:

1. Randomly 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/air quality alerts.
3. Modify the health outcome to add a treatment effect that will try to recover.
4. Estimate the model.
5. Store the point estimate of interest and its standard error.
6. Repeat the procedure 1000 times.
7. Compute the proportion of statistically significant estimates at the 5% level (the power), the average of the absolute value of significant estimates over the true effect size (the exaggeration ratio), and the proportion of significant estimates of the opposite sign of the true effect (the probability to make type S error).

Modeling assumptions. To only capture the specific issues arising due to low statistical power, we build our simulations such that (i) they meet all the required assumptions of empirical strategies and (ii) make it easier—compared to real settings—to recover the treatment effect. For all research designs, the treatment we add to the data is not biased by unmeasured confounders nor measurement errors. For instrumental variable and reduced-form strategies, we only simulate binary and randomly allocated exogenous shocks (*e.g.* the occurrence of a thermal inversion). For the regression discontinuity approach, we only model sharp designs where an air quality alert is always activated above a randomly chosen threshold. Our models always retrieve on average the true value of the treatment effect we set in the data.

Two approaches for simulating research designs. For the reduced-form and regression discontinuity designs, we follow the Neyman-Rubin causal framework by simulating all potential outcomes (Rubin 1974). We consider that the health outcome value recorded in the NMMAPS dataset corresponds to the potential outcome $Y_{c,t}(0)$. To create the counterfactuals $Y_{c,t}(1)$, we add a treatment effect drawn from a Poisson distribution whose parameter corresponds to the magnitude of the treatment. We then randomly draw the treatment indicators $T_{t,c}$ for exogenous shocks or air quality alerts. For reduced-form strategies, the treatment status of each day is drawn from a Bernoulli distribution with parameter equal to the proportion of exogenous shocks desired. For air pollution alerts, we randomly draw a threshold from a uniform distribution and select a bandwidth such that it yields the desired proportion of treated observations. We finally express the observed values Y^{obs} of potential outcomes according to the treatment assignment: $Y^{obs}_{c,t} = (1\text{-}T_{c,t})\times Y_{c,t}(0) + T_{c,t}\times Y_{c,t}(1)$.

To simulate standard regression and the instrument variable strategies, we rely on a model-based approach. For the standard regression strategy, we first estimate the following statistical model on our data:

$$Y_{c,t} = \alpha + \beta Z_{c,t} + \mathbf{W}_{c,t}\lambda + \mathbf{C}_t\gamma + \epsilon_{c,t}$$

We then predict new observations of a $Y_{c,t}$ using the estimated coefficients of the model ($\hat{\beta}$, $\hat{\lambda}$, and $\hat{\gamma}$) and by adding noise drawn from a normal distribution with variance equal to that of the residuals $\widehat{\epsilon_{c,t}}$ (Peng et al. 2006). We modify the slope of the dose-response relationship by changing the value of the air pollution coefficient β. For the instrumental variable strategy, we use the same method as for the standard regression approach but first modify observed air pollutant concentrations $P_{c,t}$ according to the desired effect size θ of the randomly allocated instrument:

$$\widetilde{P}_{c,t} = P_{c,t} + \theta Z_{c,t}$$

We draw the allocation of each day to an exogenous shock from a Bernoulli distribution with parameter equal to the proportion of exogenous shocks. We then estimate a two-stage least squares model (2SLS) and modify the coefficient for the effect of the air pollutant on an health outcome. We finally generate the fake observations of the health outcome by combining the prediction from the modified 2SLS model and noise drawn a normal distribution with variance equal to that of the residuals.

Varying parameters. To understand which parameters affect statistical power issues, we modify one aspect of the research design while keeping other parameters constant. We study the influence of four main parameters. First, we vary the sample size by drawing a different number of cities and changing the length of the study period. Second, we consider different effects size of air pollution or of an exogenous shock on the health outcome. Third, we allocate increasing proportions of exogenous shocks/air quality alerts. Fourth, we vary the number of cases in the outcome by considering different health outcomes.

4.4 Simulation of Case Studies.

The simulations described above help explore the effect of each parameter on statistical power issues. Yet, the resulting set of parameters considered may not be perfectly representative of actual studies. We therefore calibrate our simulations to reproduce three papers published in the literature.

5 Results

In this section, we first describe how statistical power evolves with the treatment effect size, the number of observations, the proportion of exogenous shocks and the average count of the health outcome. We then show that statistical power issues can be

substantial for actual parameter values found in the literature.

5.1 Evolution of Power, Exaggeration Ratio and Type S Error with Study Parameters

We analyze how statistical power, exaggeration ratio and type S error 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. We consider the following baseline parameters:

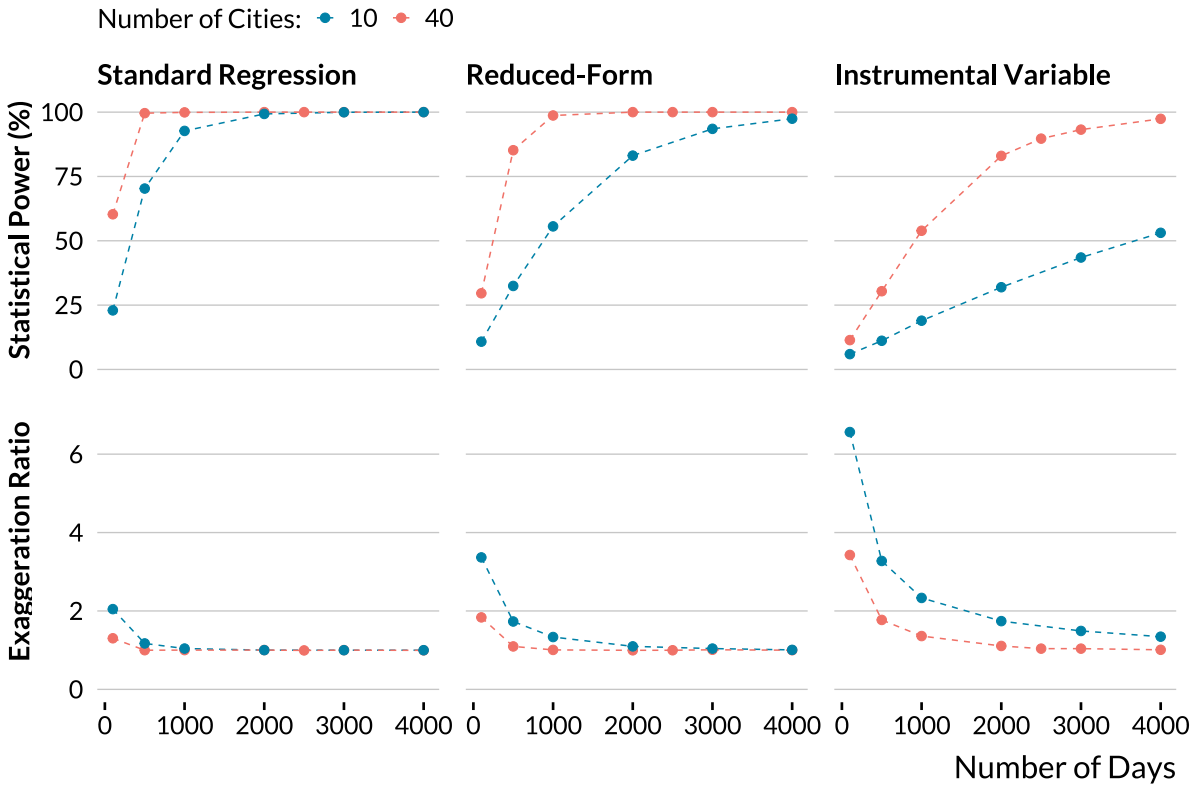
- A large sample size of 100,000 observations (2500 days × 40 cities),
- A 1% effect size, the order of magnitude found in the most precise studies of the literature. A one standard deviation in air pollution or an exogenous shock increases the health outcome by 1%,
- 50% of observations are subject of an exogenous shock. For air pollution alerts analyzed with regression discontinuity 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 (average daily mean of 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

In Figure 7, we recover the well-known increasing relationship between the number of observations and statistical power. Conversely, type-M error decreases with the number of observations.

Figure 7: Evolution of Power and Exaggeration with Sample Size.



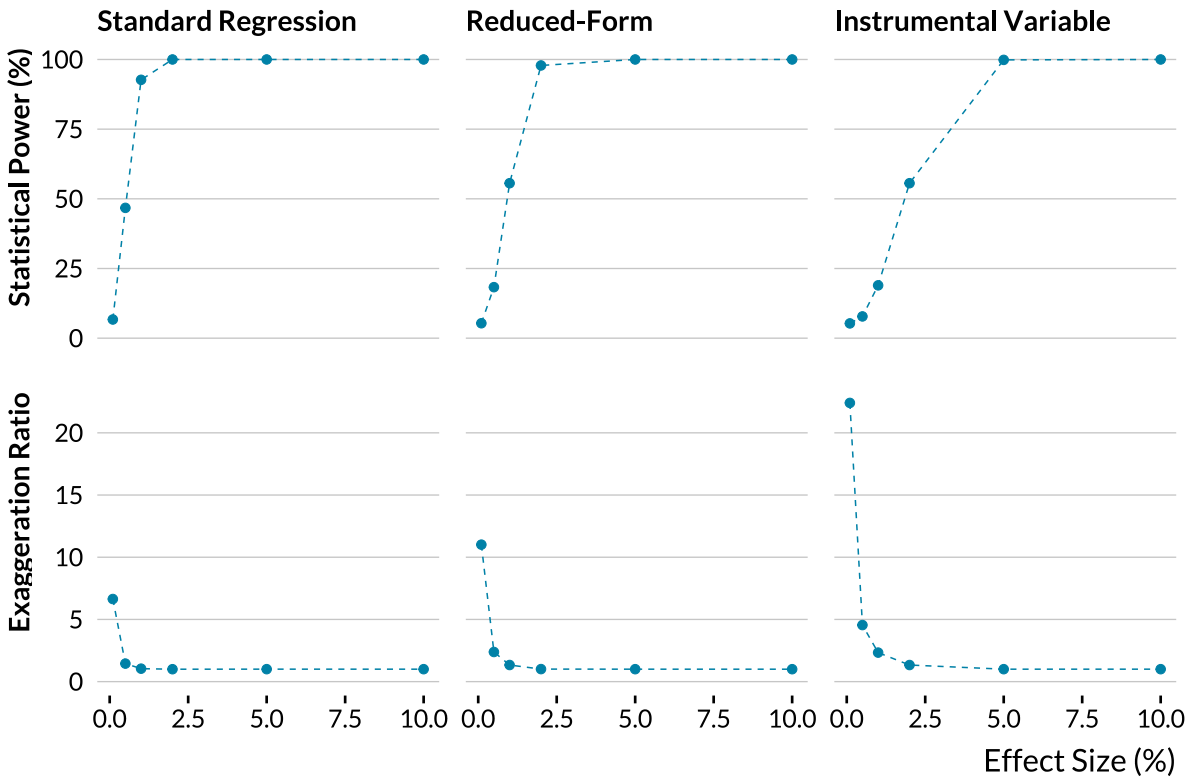
Notes: The other parameters are set to their baseline values: a true effect size of 1%, 50% of observations subject to an exogenous shock for instrumental variable and reduced-form designs, and the health outcome is the total number of non-accidental deaths.

This result comes from the fact that the variance of usual estimators decreases with the number of observations. For instance, in the homoskedastic case of the OLS, $\hat{\beta} \xrightarrow{d} \mathcal{N}(\beta, \mathbb{E}[\mathbf{X}\mathbf{X}']^{-1}\sigma^2/n)$, where n is the number of observations, $\hat{\beta}$ the OLS estimate of β, the parameter of interest associated with $\mathbf{X}$. As the variance decreases, the statistical power increases. As the variance decreases, the statistical power increases and exaggeration decreases (Zwet and Cator 2021, Lu et al. 2019).

We also find that statistical power and exaggeration issues can arise even for a large number of observations. For a sample size of 40,000 observations, the instrumental variable strategy only has a statistical power of 54% and overestimates the true effect by a factor of 1.4. On the contrary, the 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, the probability to make a Type S error is null For all identification methods and sample sizes.

Effect Size

Figure 8: Evolution of Power and Exaggeration with Effect Size.



Notes: The other parameters are set to their baseline values: a sample size of 100,000, 50% of observations subject to an exogenous shock for instrumental variable and reduced-form designs, and the health outcome is the total number of non-accidental deaths.

In Figure 8, we retrieve another familiar result: the larger the effect size, the larger

the power. As expected based on Zwet and Cator (2021) and Lu et al. (2019)'s results, we also find that the exaggeration ratio decreases with the true effect size. Even for our large baseline sample size, statistical power issues appear for effect sizes routinely found in the epidemiology literature. For instance, for our instrumental variable strategy and an effect size of 0.5%, the average exaggeration ratio is about 1.7. As for results on sample sizes, standard regression and reduced-form strategies are less prone to power issues, even for small effect sizes.

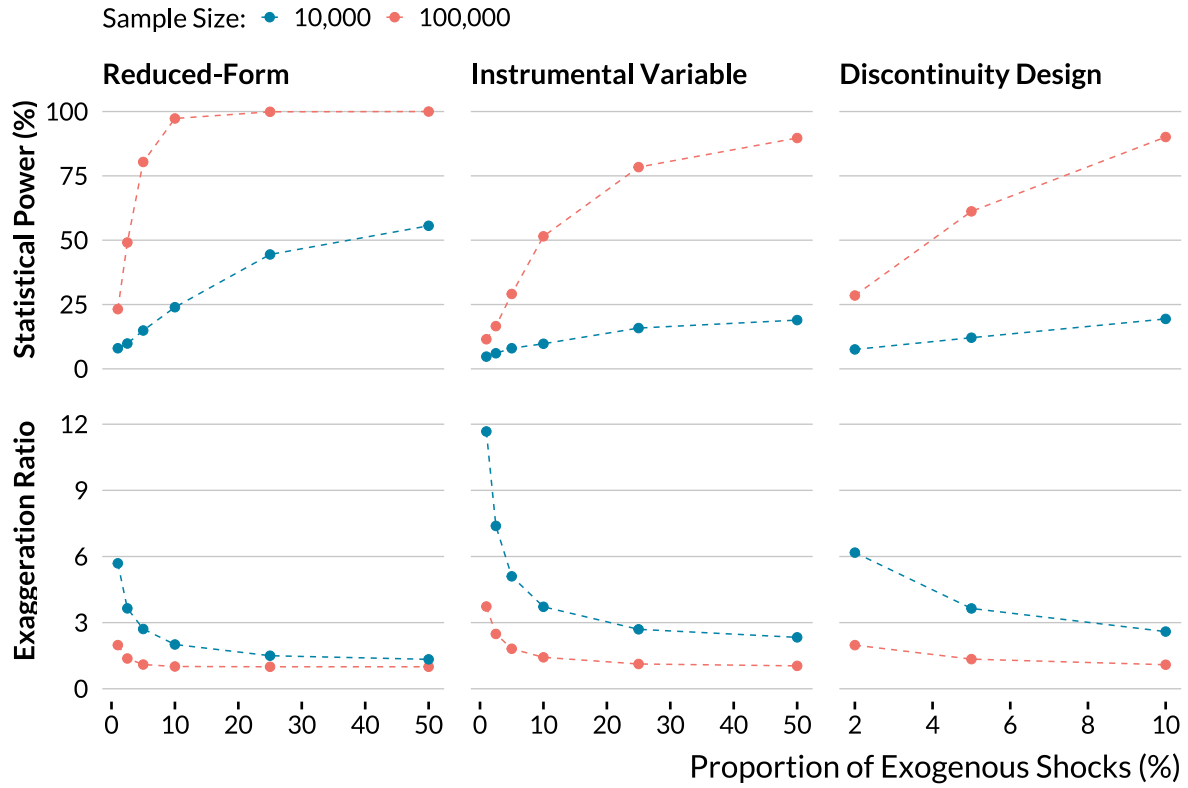
Proportion of Exogenous Shocks

The link between the proportion of exogenous shocks and statistical power might be less widely known. In Figure 9, we show that statistical power increases with the proportion of treated units for instrumental variable, regression discontinuity and reduced-form designs. Conversely, the average exaggeration ratio increases as the proportion of exogenous shocks decreases.

As in the case of randomized controlled trials, the precision of studies will be maximized when half of the observations are exposed to the treatment of interest. The variance of the average treatment effect estimator (ATE) is $\sigma^2/[n \times p(1-p)]$ where σ is the standard deviation of the outcome in the treated and control groups and p the proportion of treated units. This variance increases when p departs from 0.5. Another way to interpret this result is to consider that a small number of exogenous shocks limits the variation that can be leverage to identify the effect of interest. When the proportion of shocks decreases, the variance of the treatment variable decreases and therefore the variance of the estimator increases. A similar reasoning can be applied to IV strategies.

In practice, air pollution alerts, thermal inversion or transportation strikes are generally rare events. In some studies, they represent less than 5% of the observations. With a dataset of 10,000 observations, our simulations return an average exaggeration

Figure 9: Evolution of Power and Exaggeration with the Proportion of Exogenous Shocks.



Notes: The other parameters are set to their baseline values: a true effect size of 1% and the health outcome is the total number of non-accidental deaths. The proportion of exogenous shocks corresponds to the fraction of days in the sample that are allocated to the treatment.

ratio of 2.7 for the reduced-form strategy. Despite large sample sizes, air pollution studies exploiting few exogenous shocks might be particularly prone to exaggeration issues.

Average Count of Cases of the Health Outcome

Subgroup analyses are routinely carried out in the literature to evaluate the acute health effects of air pollution on children or the elderly. Yet, the average count of cases can also critically affect statistical power as shown in Table 2. For instance, in a setting with only a few deaths per day, a 1% increase in the number of deaths will rarely cause additional deaths. The effect will be more difficult to detect. To simulate

Table 2: Evolution of Power and Exaggeration with the Average Number of Daily Cases of Health Outcomes.

	Non-Accidental	Respiratory	COPD
Number of Cases	23	2	0.3
Statistical Power (%)	90	16	7.5
Exaggeration Ratio	1	2.4	5.9

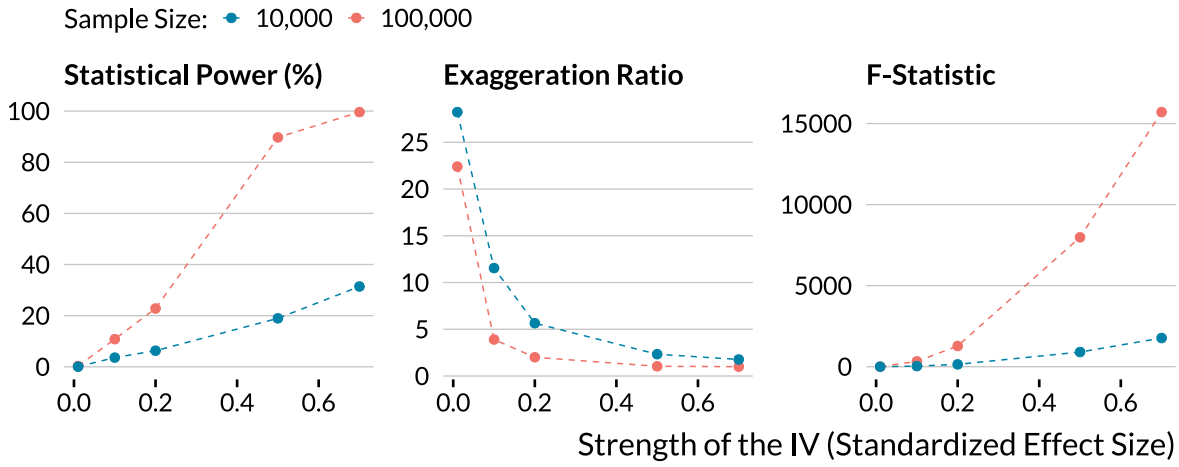
Notes: This table displays the average number of cases, the power and the exaggeration ratio for three health outcomes: non-accidental deaths, respiratory deaths, and chronic pulmonary deaths for individuals aged between 65 and 75. These figures are obtained for the instrumental variable design with a sample size of 100,000 and 50% of observations subject to an exogenous shock. The instrument variable increases the air pollutant concentration by 0.5 standard deviation. A one standard deviation increase in the instrumented air pollutant leads to 1% relative increase in the health outcome considered.

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 (COPD) cases for people aged between 65 and 75 (daily mean $\simeq$ 0.3). Using baseline parameters and in the case of the large dataset, we find that statistical power is close to 100% when for a 1% increase in the total number of non-accidental deaths. However, statistical power drops when the average count of cases decreases. For instance, the instrumental variable strategy has only 16% of statistical power to detect a 1% increase in respiratory deaths. The average exaggeration ratio is then equal to 2.4. For chronic obstructive pulmonary deaths—the health outcome with lowest number of cases—the situation is even worst, the average exaggeration ratio reaches 5.9. When focusing on subgroups such as children or the elderly, one can expect to find larger effect sizes as those populations are more vulnerable to air pollution. While these larger effect sizes attenuate exaggeration concerns, the lower number of cases exacerbates them. It creates a trade-off for power issues.

Issues Specific to the Instrumental Variable Design

In the case of instrumental variable strategies, statistical power is affected by the strength of the instrument. In our simulations, we consider a binary instrument (e.g., the occurrence of a thermal inversion or a public transport strike). We define its strength as the standardized effect size of the instrument on the air pollutant concentration. A strength of 0.2 means that the instrument increases the concentration by 0.2 standard deviation.

Figure 10: Evolution of Power and Exaggeration with the Strength of the Instrumental Variable.



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 strength of the instrumental variable is defined as its effect in standard deviation on the air pollutant concentration.

As shown in Figure 10, we find that statistical power collapses and exaggeration soars when the instrument's strength decreases. Importantly, this issue even arises for large first-stage F-statistics. In our simulations based the large data set with 100,000 observations, an instrumental variable's strength of 0.2, and an effect size of 1%, we find an average F-statistics of 1278. The statistical power is however only 23% and the average exaggeration ratio 2. A large F-statistic could therefore hide a weak instrumental variable that results in a low statistical power and large exaggeration.

The relationship between IV strength and exaggeration comes from the fact that the variance of the 2SLS estimator decreases with the correlation between the instrument and the instrumented variable. In the homoskedastic case, the asymptotic variance of the 2SLS estimator is $\left(\mathbb{E}[XZ']\mathbb{E}[ZZ']^{-1}\mathbb{E}[ZX']\right)^{-1}\sigma^2$, where σ^2 is variance of the error, X the endogenous variable and Z the instrument. When $\mathbb{E}[XZ']$ and $\mathbb{E}[ZX']$ decrease, the variance of the estimator increases. Zwet and Cator (2021) and Lu et al. (2019) show that as the variance of a normally distributed estimator increases, the statistical power decreases and exaggeration increases.

5.2 Case Studies

The previous simulation results help understand how the various parameters influence the statistical power of studies. Yet, these parameters may not perfectly represent actual studies as we made several conservative assumptions: relatively large sample size, proportion of treated units, average outcome counts and instrumental variable strength. For each research design, we therefore consider a realistic set of parameters based on an example from the literature. We then vary the value of key parameters. As we are working with different data, we cannot exactly reproduce the level of precision found in the articles considered. Our goal is not to claim that the estimates produced by a particular article are inflated, but instead to understand how low power issues could arise for representative parameter values.

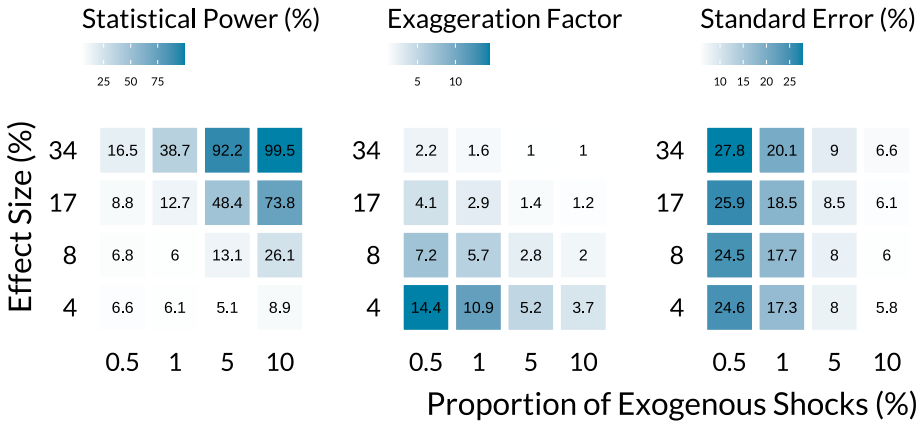
Public Transportation Strikes

Public transportation strikes are unique but rare positive shocks to air pollution as individuals use their cars to reach city centers. Even in a large data set, with several cities and a long study period, the proportion of affected days might be very small. For instance, Bauernschuster et al. (2017) investigate the effect of public transportation strikes on air pollution and emergency admission in the five biggest German cities

over a period of 6 years. Despite a sample size of 11,000, there are only 57 1-day strikes during the study period (0.5% of days are actually treated). The authors find that children hospitalizations for breathing issues increase by 34% (SE=8%) on strike days. On average, 0.22 children per day go to the hospital for breathing issues.

We simulate a similar design with our own data. We first randomly sample 2200 observations for five cities and then vary (i) the proportion of exogenous shocks from 0.5% up to 10%, and (ii) the treatment effect size from a 4% increase up to a 34% increase. We focus on elderly mortality due to chronic obstructive pulmonary disease since it has an average daily count of 0.29 cases.

Figure 11: Evolution of Power and Exaggeration for Public Transportation Strikes Designs.



Notes: Each panel displays the average value of a metric (power, exaggeration, and standard error) for varying proportions of exogenous shocks and effect sizes. The average standard error of simulations is the raw standard error divided by the mean number of cases of the health outcome. For each combination of parameters, we ran 1000 simulations.

In Figure 11, we display our simulation results. The first panel from the left shows that both large effect sizes and a large proportion of exogenous shocks are required to reach adequate power. In the middle panel, we show that a proportion of 0.5% of exogenous shocks is associated with very large exaggeration ratios, from 2.2 for a true effect size of 34% up to 14 for one of 4%. Power issues fade for a combination of a proportion of exogenous shocks above 5% and effect sizes above 17%. In the right

panel, we plot the average standard error of the estimates, expressed as a fraction of the average of the health outcome. The standard error of Bauernschuster et al. (2017)'s is 8%. In our simulations, we recover that specific precision for a proportion of exogenous shocks of 5%. In that case, a true effect size of 34% would not yield inflated estimates. However, if effect sizes are actually smaller and more representative of those found in the literature, the exaggeration would be consequential.

This simulation exercise shows that exaggeration is likely to arise in practice since the proportion of exogenous shocks is low. It occurs even when true effect sizes are relatively large.

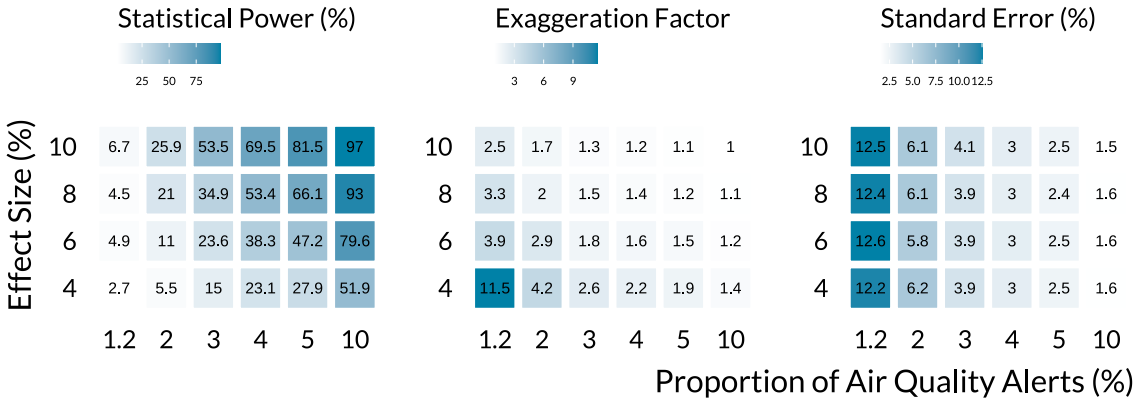
Air Pollution Alerts

Air pollution alerts are also rare events. Their effects are estimated using regression discontinuity designs that restrict the analysis to observations closed to the air quality threshold. As a consequence, the effective sample size may be particularly small. For instance, Chen et al. (2018) investigate the effects of air quality alerts on emergency department visits in Torento, over the 2003-2012 period. While the nominal sample size is 3652, the effective one is only 143 (100 control days and 43 treated days). Only 1.2% of observations are treated. The authors find that eligibility to air quality (the intention-to-treat effect) approximately reduces emergency visits for asthma by 8% (SE=3.8%). The average daily count of cases of their health outcome is 26.

We approximate the setting of Chen et al. (2018) using our data. We first sample one city for a time period of 3652 days and randomly allocate the treatment. We then repeat the process varying the proportion of alerts and effect sizes. Our outcome variable is the total number of non-accidental deaths since it has a daily mean of 23.

Figure 12 displays the simulations results. As in Figure 11, a combination of large effect sizes and many air quality alerts is needed to avoid low power issues. We get a precision similar to Chen et al. (2018) for a proportion of air quality alerts of 3%.

Figure 12: Evolution of Power and Exaggeration for Air Quality Alerts Designs.



Notes: Each panel display the average value of a metric (power, exaggeration, and standard error) for varying proportions of exogenous shocks and effect sizes. The average standard error of simulations is the raw standard error divided by the mean number of cases of the health outcome. For each combination of parameters, we ran 1000 simulations.

For an effect size of 4%, the average exaggeration ratio is equal to 2.6. In that case, the average average of statistically significant estimates is 10%, which is similar to the effect size found by Chen et al. (2018).

Unless true effect sizes are very large, air quality alert designs produce inflated estimates in realistic settings.

Instrumenting Air Pollution

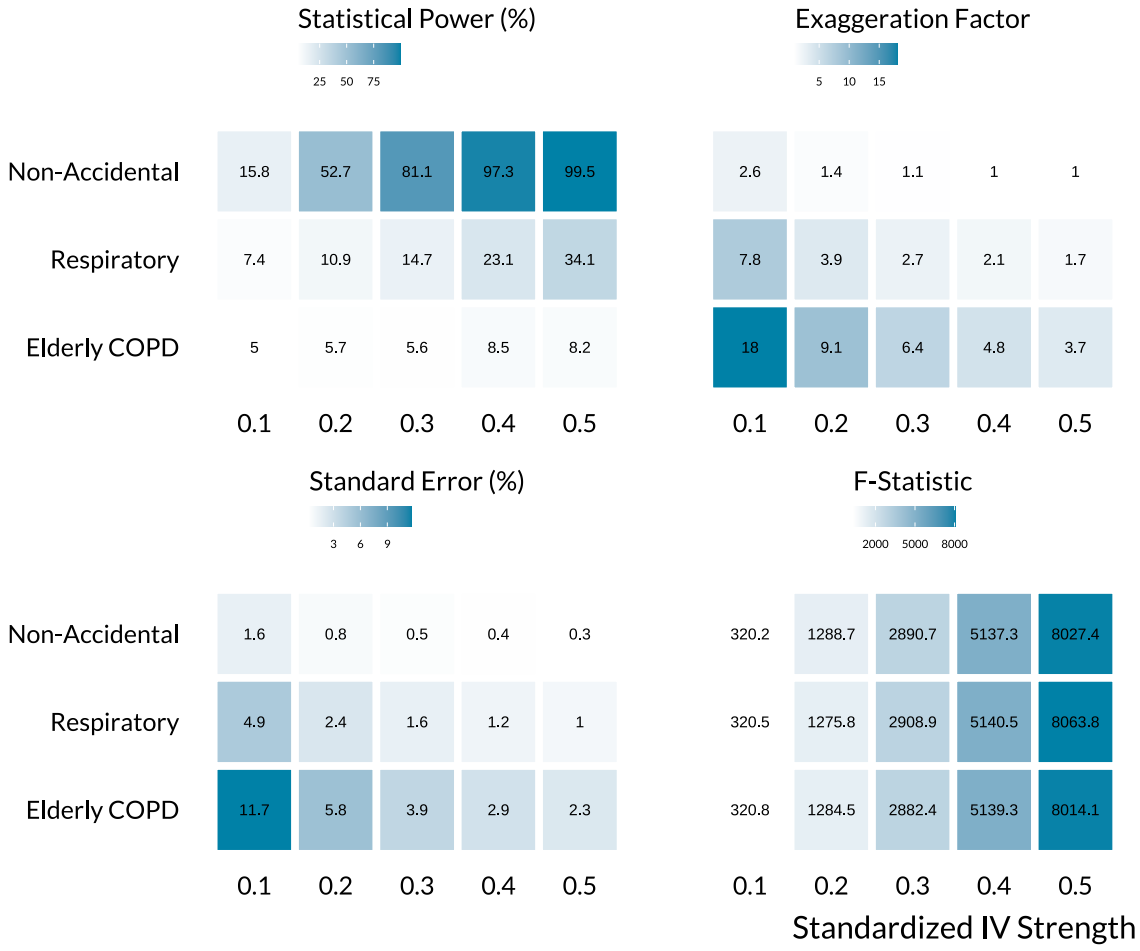
Finally, we investigate the most commonly used strategy in the causal inference literature, the instrumental variable design. Several studies rely on very large datasets and exploit changes in weather patterns as sources of exogenous variations. For instance, Schwartz et al. (2018) instrument $PM_{2.5}$ concentration with planetary boundary layer, winds speed, and air pressure. Once the effects of seasonal and other weather parameters are accounted for, the combination of their instruments explains 18% of the variation in $PM_{2.5}$ concentration. They find that a 10 $\mu g/m^3$ increase in $PM_{2.5}$ leads to a 1.5% (SE=0.22%) increase in daily non-accidental mortality. There are on average 23 daily deaths in their dataset of 591,570 observations (135 cities with a length of study

of approximately 4382 days).

In our simulations, we assess how the strength of the instrumental variable affects power issues for several health outcomes. We consider a binary instrumental variable and vary its effect on air pollution concentration from a 0.1 to a 0.5 standard deviation increase. The 18% correlation in Schwartz et al. (2018) corresponds to a 0.4 standard deviation increase in our case (Lipsey and Wilson 2001). We assume that half of the observations are exposed to exogenous shocks. We set an effect size corresponding to a 1.5% relative increase in three health outcomes with different average number of cases: non-accidental mortality (mean cases of 23), respiratory mortality (mean of 2), and chronic obstructive pulmonary mortality of elderly (mean of 0.3). Our data set being smaller than the one used in Schwartz et al. (2018), we only run simulations for a sample size of 100,000.

In Figure 13, we see in the top-left panel that power reaches satisfactory level for large instrumental variable strengths but only for non-accidental causes. For respiratory and elderly mortality, exaggeration can be substantial even for large IV strength. While our sample size is large, it is smaller than the one in Schwartz et al. (2018). As a consequence, our simulations only have a precision close to theirs for an instrumental variable strength of 0.5 and non-accidental mortality. Yet, our simulations highlight that important exaggeration issues can arise in realistic settings, even for large IV strength. The bottom-right panel of Figure 13 confirms the result found in the simulations of the previous section: a large first stage F-statistic can be a poor indicator of statistical power issues. For instance, for non-accidental mortality and an IV strength of 0.1, the F-statistic is equal to 320 but the exaggeration factor is 2.6, with an associated power is 16%. Importantly, as the F-statistic does not vary with the number of cases in the outcome it can all the more hide important power issues.

Figure 13: Evolution of Power and Exaggeration for Instrumental Variable Designs.



Notes: Each panel display the average value of a metric (power, exaggeration, standard error, and first-stage F-statistic.) for varying proportions of exogenous shocks and effect sizes. The average standard error of simulations is the raw standard error divided by the mean number of cases of the health outcome. For each combination of parameters, we ran 1000 simulations.

6 Discussion

Growing evidence shows that a large share of results published in economics might be exaggerated due to low statistical power and a publication bias towards statistical significance (?Ioannidis et al. 2017, Brodeur et al. 2020, Ferraro and Shukla 2020). Although this issue appears to be increasingly acknowledged, guidance to tackle it is

still lacking in applied economic research (Altoè et al. 2020, Black et al. 2022). In this paper, we fill this gap by implementing a principled workflow to assess if and understand why estimates published in a particular literature could be inflated.

Our retrospective analysis showed that most studies in the standard epidemiology and causal inference literatures are not likely to suffer from substantial exaggeration issues. However, about a quarter of published results may be inflated. As a robustness exercise when carrying out a study, we recommend to systematically run retrospective calculations to gauge the risk of exaggeration. They are easy to implement and force us to evaluate if our research design enables us to confidently estimate a credible range of effect sizes. In Appendix A.2, we illustrate this approach by considering the example of Deryugina et al. (2019). Yet, a retrospective analysis does not help understand which parameters of the research design influence statistical power and exaggeration.

Our prospective analysis based on real-data simulations overcomes this limitation and leads to issue four warnings. First, sample size matters for all causal inference methods. It is particularly problematic in the case of regression-discontinuity designs applied to air pollution alerts. Second, reduced-form analyses exploiting exogenous shocks such as transport strikes often rely on too few events and can therefore produce inflated effects. Third, although it is well-known that two-stage least square estimates are inherently less precise than ordinary least squares, we show that it also makes instrumental variable strategies more prone to exaggeration issues. In cases where omitted variable and attenuation biases might be of little concern, the benefits of using an instrumental variable strategy could be questioned. In a companion paper, we explore the trade-off between targeting an unbiased estimate with causal inference methods and exaggerating effect sizes due to low power issues (Bagilet and Zabrocki-Hallak 2022). In that paper, we show how tools such as quantitative bias analyses can help position ourselves with respect to this trade-off (Oster 2019, Rosenbaum 2020, Cinelli and Hazlett 2020). Fourth, for all research designs, exaggeration 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 an actual risk of exaggeration, even for large sample sizes.

These results highlight the importance of implementing prospective simulations before running an observational study. Fake-data can be simulated from scratch or simulations can build on datasets used in other studies. In our replication material, we provide a template to run such simulations to ease the adoption of this practice.

On top of these specific recommendations, we should not forget that published estimates only suffer from exaggeration in the presence of publication bias. The causal inference literature would therefore benefit from adopting a different view towards statistically insignificant results (Ziliak and McCloskey 2008, Wasserstein and Lazar 2016, McShane et al. 2019). It currently dichotomizes evidence according to the 5% significance threshold, disregarding non-significant results (Greenland 2017). Instead, if results were published regardless of their significance, the resulting distribution would be centered around the true effect (Hernán 2022). To replace the null hypothesis testing framework, we recommend to focus on confidence intervals and to interpret the range of effect sizes supported by the data (Amrhein et al. 2019, Romer 2020).

Qualifying estimates as "statistically significant" does not acknowledge the actual uncertainty that should be computed and embraced to better help policy-makers evaluate the adverse effects of air pollution. Prospective and retrospective power analyses can help design better studies and improve the interpretation of their results.

References

Adhvaryu, Achyuta, Namrata Kala, and Anant Nyshadham (2022) "Management and shocks to worker productivity," *Journal of Political Economy*, 130 (1), 1–47.

Altoè, Gianmarco, Giulia Bertoldo, Claudio Zandonella Callegher, Enrico Toffalini,

Antonio Calcagnì, Livio Finos, and Massimiliano Pastore (2020) "Enhancing Statistical Inference in Psychological Research via Prospective and Retrospective Design Analysis," *Frontiers in Psychology*, 10, 2893, 10.3389/fpsyg.2019.02893.

Amrhein, Valentin, David Trafimow, and Sander Greenland (2019) "Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication," *The American Statistician*, 73 (sup1), 262–270, 10.1080/00031305.2018.1543137.

Anderson, Michael L, Minwoo Hyun, and Jaecheol Lee (2022) "Bounds, Benefits, and Bad Air: Welfare Impacts of Pollution Alerts,"Technical report, National Bureau of Economic Research.

Angrist, Joshua D. and Guido W. Imbens (1995) "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90 (430), 431–442, 10.1080/01621459.1995.10476535.

Arceo, Eva, Rema Hanna, and Paulina Oliva (2016) "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, 10.1111/ecoj.12273.

Arel-Bundock, Vincent, Ryan C Briggs, Hristos Doucouliagos, Marco Mendoza Aviña, and T.D. Stanley (2022) "Quantitative Political Science Research is Greatly Underpowered," preprint, Open Science Framework, 10.31219/osf.io/7vy2f.

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

Bagilet, Vincent and Léo Zabrocki-Hallak (2022) "Uncounfounded But Inflated Causal Estimates," *CEEP Working Paper Series* (20), https://ceep.columbia.edu/sites/default/files/content/papers/n20.pdf.

Barwick, Panle Jia, Shanjun Li, Deyu Rao, and Nahim Bin Zahur (2018) "The Morbidity Cost of Air Pollution: Evidence from Consumer Spending in China,"Technical Report w24688, National Bureau of Economic Research, Cambridge, MA, 10.3386/w24688.

Bauernschuster, Stefan, Timo Hener, and Helmut Rainer (2017) "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, 10.1257/pol.20150414.

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

Bind, Marie-Abèle (2019) "Causal Modeling in Environmental Health," *Annual Review of Public Health*, 40 (1), 23–43, 10.1146/annurev-publhealth-040218-044048.

Black, Bernard, Alex Hollingsworth, Letícia Nunes, and Kosali Simon (2022) "Simulated power analyses for observational studies: An application to the Affordable Care Act Medicaid expansion," *Journal of Public Economics*, 213, 104713, 10.1016/j.jpubeco.2022.104713.

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

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg (2016) "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics*, 8 (1), 1–32, 10.1257/app.20150044.

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò (2013) "Power failure: why small sample size undermines the reliability of neuroscience," *Nature Reviews Neuroscience*, 14 (5), 365–376, 10.1038/nrn3475.

Camerer, Colin F., Anna Dreber, Felix Holzmeister et al. (2018) "Evaluating the replica-

bility of social science experiments in Nature and Science between 2010 and 2015," *Nature Human Behaviour*, 2 (9), 637–644, 10.1038/s41562-018-0399-z.

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

Chen, Siyu, Chongshan Guo, and Xinfei Huang (2018) "Air Pollution, Student Health, and School Absences: Evidence from China," *Journal of Environmental Economics and Management*, 92, 465–497, 10.1016/j.jeem.2018.10.002.

Cheung, Chun Wai, Guojun He, and Yuhang Pan (2020) "Mitigating the air pollution effect? The remarkable decline in the pollution-mortality relationship in Hong Kong," *Journal of Environmental Economics and Management*, 101, 102316, 10.1016/j.jeem.2020.102316.

Cinelli, Carlos and Chad Hazlett (2020) "Making sense of sensitivity: extending omitted variable bias," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82 (1), 39–67, 10.1111/rssb.12348.

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

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

Dominici, Francesca and Corwin Zigler (2017) "Best Practices for Gauging Evidence of Causality in Air Pollution Epidemiology," *American Journal of Epidemiology*, 186 (12), 1303–1309, 10.1093/aje/kwx307.

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

Ebenstein, Avraham, Victor Lavy, and Sefi Roth (2016) "The Long-Run Economic Consequences of High-Stakes Examinations: Evidence from Transitory Variation in Pollution," *American Economic Journal: Applied Economics*, 8 (4), 36–65, 10.1257/app.20150213.

Fan, Maoyong, Guojun He, and Maigeng Zhou (2020) "The winter choke: Coal-Fired heating, air pollution, and mortality in China," *Journal of Health Economics*, 71, 102316, 10.1016/j.jhealeco.2020.102316.

Fan, Maoyong and Yi Wang (2020) "The impact of PM2.5 on mortality in older adults: evidence from retirement of coal-fired power plants in the United States," *Environmental Health*, 19 (1), 28, 10.1186/s12940-020-00573-2.

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

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

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

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

Giaccherini, Matilde, Joanna Kopinska, and Alessandro Palma (2021) "When particulate matter strikes cities: Social disparities and health costs of air pollution," *Journal of Health Economics*, 78, 102478, 10.1016/j.jhealeco.2021.102478.

Godzinski, Alexandre, M Suarez Castillo et al. (2019) "Short-term health effects of pub-

lic transport disruptions: air pollution and viral spread channels,"Technical report, Institut National de la Statistique et des Etudes Economiques.

Godzinski, Alexandre and Milena Suarez Castillo (2021) "Disentangling the effects of air pollutants with many instruments," *Journal of Environmental Economics and Management*, 109, 102489, 10.1016/j.jeem.2021.102489.

Greenland, Sander (2017) "Invited Commentary: The Need for Cognitive Science in Methodology," *American Journal of Epidemiology*, 186 (6), 639–645, 10.1093/aje/kwx259.

Griffin, Beth Ann, Megan S. Schuler, Elizabeth A. Stuart et al. (2021) "Moving beyond the classic difference-in-differences model: a simulation study comparing statistical methods for estimating effectiveness of state-level policies," *BMC Medical Research Methodology*, 21 (1), 279, 10.1186/s12874-021-01471-y.

Guidetti, Bruna, Paula Pereda, and Edson Severnini (2021) ""Placebo Tests" for the Impacts of Air Pollution on Health: The Challenge of Limited Health Care Infrastructure," *AEA Papers and Proceedings*, 111, 371–375, 10.1257/pandp.20211031.

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

Hanlon, W Walker (2018) "London Fog: A Century of Pollution and Mortality, 1866-1965," *The Review of Economics and Statistics*, 1–49.

He, Guojun, Maoyong Fan, and Maigeng Zhou (2016) "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, 10.1016/j.jeem.2016.04.004.

He, Guojun, Tong Liu, and Maigeng Zhou (2020) "Straw burning, PM2.5, and death: Evidence from China," *Journal of Development Economics*, 145, 102468, 10.1016/j.jdeveco.2020.102468.

Hernán, Miguel A. (2022) "Causal analyses of existing databases: no power calcula-

tions required," *Journal of Clinical Epidemiology*, 144, 203–205, 10.1016/j.jclinepi.2021.08.028.

Herrnstadt, Evan, Anthony Heyes, Erich Muehlegger, and Soodeh Saberian (2021) "Air Pollution and Criminal Activity: Microgeographic Evidence from Chicago," *American Economic Journal: Applied Economics*, 13 (4), 70–100, 10.1257/app.20190091.

Ioannidis, John P. A. (2008) "Why Most Discovered True Associations Are Inflated," *Epidemiology*, 19 (5), 640–648, http://www.jstor.org/stable/25662607.

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

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

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

Kim, Moon Joon (2021) "Air Pollution, Health, and Avoidance Behavior: Evidence from South Korea," *Environmental and Resource Economics*, 79 (1), 63–91, 10.1007/s10640-021-00553-1.

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

Le Tertre, A, S Medina, E Samoli et al. (2002) "Short-term effects of particulate air pollution on cardiovascular diseases in eight European cities," *Journal of Epidemiology & Community Health*, 56 (10), 773–779.

Lipsey, Mark W and David B Wilson (2001) *Practical meta-analysis.*: SAGE publications, Inc.

Liu, Cong, Renjie Chen, Francesco Sera et al. (2019) "Ambient Particulate Air Pollution and Daily Mortality in 652 Cities," *New England Journal of Medicine*, 381 (8), 705–715, 10.1056/NEJMoa1817364.

Liu, Ya-Ming and Chon-Kit Ao (2021) "Effect of air pollution on health care expenditure: Evidence from respiratory diseases," *Health Economics*, 30 (4), 858–875, 10.1002/hec.4221.

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

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

McShane, Blakeley B., David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett (2019) "Abandon Statistical Significance," *The American Statistician*, 73 (sup1), 235–245, 10.1080/00031305.2018.1527253.

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

Mullins, Jamie and Prashant Bharadwaj (2015) "Effects of Short-Term Measures to Curb Air Pollution: Evidence from Santiago, Chile," *American Journal of Agricultural Economics*, 97 (4), 1107–1134, 10.1093/ajae/aau081.

Open Science Collaboration (2015) "Estimating the reproducibility of psychological science," *Science*, 349 (6251), aac4716, 10.1126/science.aac4716.

Oster, Emily (2019) "Unobservable selection and coefficient stability: Theory and evidence," *Journal of Business & Economic Statistics*, 37 (2), 187–204.

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

Peng, Roger D, Francesca Dominici, and Thomas A Louis (2006) "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.

Romer, David (2020) "In praise of confidence intervals," in *AEA Papers and Proceedings*, 110, 55–60.

Rosenbaum, Paul R. (2020) *Design of Observational Studies*, Springer Series in Statistics, Cham: Springer International Publishing, 10.1007/978-3-030-46405-9.

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

Samet, Jonathan M, Scott L Zeger, Francesca Dominici, Frank Curriero, Ivan Coursac, Douglas W Dockery, Joel Schwartz, and Antonella Zanobetti (2000) "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.

Schell, Terry, Beth Ann Griffin, and Andrew Morral (2018) *Evaluating Methods to Estimate the Effect of State Laws on Firearm Deaths: A Simulation Study*: RAND Corporation, 10.7249/RR2685.

Schlenker, Wolfram and W. Reed Walker (2016) "Airports, Air Pollution, and Contemporaneous Health," *The Review of Economic Studies*, 83 (2), 768–809, 10.1093/restud/rdv043.

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

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

Schwartz, Joel, Marie-Abele Bind, and Petros Koutrakis (2017) "Estimating Causal Effects of Local Air Pollution on Daily Deaths: Effect of Low Levels," *Environmental*

Health Perspectives, 125 (1), 23–29, 10.1289/EHP232.

Schwartz, Joel, Kelvin Fong, and Antonella Zanobetti (2018) "A National Multicity Analysis of the Causal Effect of Local Pollution, NO2, and PM2.5 on Mortality," *Environmental Health Perspectives*, 126 (8), 087004, 10.1289/EHP2732.

Shah, Anoop S V, Kuan Ken Lee, David A McAllister et al. (2015) "Short term exposure to air pollution and stroke: systematic review and meta-analysis," *BMJ*, h1295, 10.1136/bmj.h1295.

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

Simeonova, Emilia, Janet Currie, Peter Nilsson, and Reed Walker (2021) "Congestion Pricing, Air Pollution, and Children's Health," *Journal of Human Resources*, 56 (4), 971–996, 10.3368/jhr.56.4.0218-9363R2.

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

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

Vichit-Vadakan, Nuntavarn, Nitaya Vajanapoom, and Bart Ostro (2008) "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, 10.1289/ehp.10849.

Wasserstein, Ronald L. and Nicole A. Lazar (2016) "The ASA Statement on p -Values: Context, Process, and Purpose," *The American Statistician*, 70 (2), 129–133, 10.1080/00031305.2016.1154108.

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

Xia, Fan, Jianwei Xing, Jintao Xu, and Xiaochuan Pan (2022) "The short-term impact of air pollution on medical expenditures: Evidence from Beijing," *Journal of Environmental Economics and Management*, 114, 102680, 10.1016/j.jeem.2022.102680.

Zhong, Nan, Jing Cao, and Yuzhu Wang (2017) "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, 10.1086/692115.

Ziliak, Stephen Thomas and Deirdre N. McCloskey (2008) *The cult of statistical significance: how the standard error costs us jobs, justice, and lives*, Economics, cognition, and society, Ann Arbor: University of Michigan Press, OCLC: ocn168717577.

Zwet, Erik W. and Eric A. Cator (2021) "The significance filter, the winner's curse and the need to shrink," *Statistica Neerlandica*, 75 (4), 437–452, 10.1111/stan.12241.

A Appendix

A.1 List of Studies Included in the Causal Inference Literature

We display below studies included in the retrospective analysis of the causal inference literature. We group them by research designs:

Instrumental Variable Design: Moretti and Neidell (2011), Ebenstein et al. (2015), Schwartz et al. (2015), Arceo et al. (2016), He et al. (2016), Knittel et al. (2016), Schlenker and Walker (2016), Sheldon and Sankaran (2017), Schwartz et al. (2017), Zhong et al. (2017), Barwick et al. (2018), Hanlon (2018), Schwartz et al. (2018), Halliday et al. (2019), Deryugina et al. (2019), Cheung et al. (2020), Fan and Wang (2020), He et al. (2020), Giaccherini et al. (2021), Godzinski and Suarez Castillo (2021), Guidetti et al. (2021), Kim (2021), Liu and Ao (2021), Xia et al. (2022)

Reduced-Form Design: Bauernschuster et al. (2017), Jans et al. (2018), Jia and Ku (2019), Godzinski et al. (2019)

Regression Discontinuity Design: Chen et al. (2018), Fan et al. (2020), Anderson et al. (2022)

Event-Study Design: Mullins and Bharadwaj (2015), Simeonova et al. (2021)

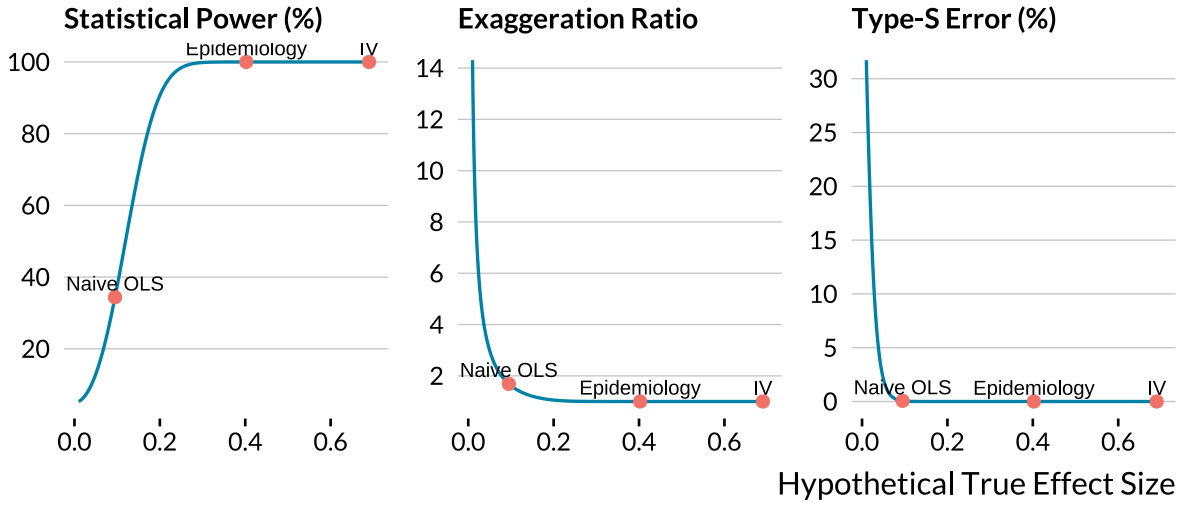
Matching Design: Baccini et al. (2017), Forastiere et al. (2020)

A.2 Implementing a Retrospective Power Analysis

We explain here how we can easily implement a retrospective power analysis once a study is completed. 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 A.1, 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 A.1: 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 " Naive 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 A.1. 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 larger than this estimate labeled as "Epidemiology" in Figure A.1. 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 could be null but the overestimation 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 naive ordinary least square estimate. Yet, the authors could argue that their instrumental variable strategy leads to a higher effect size as it overcomes unmeasured counfounding bias and measurement error. 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.