Air Quality and Health Effects

Keeping Count of Government Science:

P-Value Plotting, P-Hacking, and $PM_{2.5}$ Regulation

S. Stanley Young
Warren Kindzierski
David Randall

Our study

Far too frequently scientists cannot replicate claims made in research.

Our study examined research in the field of epidemiology that informs US EPA regulation of PM2.5 in ambient air.

Used *counting* & *p-value plots* to independently test meta-analysis studies making PM2.5– mortality, heart attack & asthma claims.

Allows us to judge whether US EPA regulations on PM2.5 reflect irreproducible, flawed or unsound research.

We used 2 methods – *counts of statistical tests* & *p-value plots* – to provide independent (severe) tests of meta-analyses.

- Counting involves estimating the number of statistical tests performed in a base study. A base study is a study selected for quantitative evaluation in meta-analysis.
- We converted risk statistics from the base studies – Relative Risks or Odds Ratios with confidence intervals – to p-values and then we ordered them smallest to largest and plotted them in a p-value plot. This allowed us to examine the nature of the distribution of statistics combined in meta-analysis – we will show you shortly what a p-value plot looks like.
- This allowed us to judge whether US EPA regulations on PM2.5 reflect unsound research.

Multiple testing & multiple modeling problem

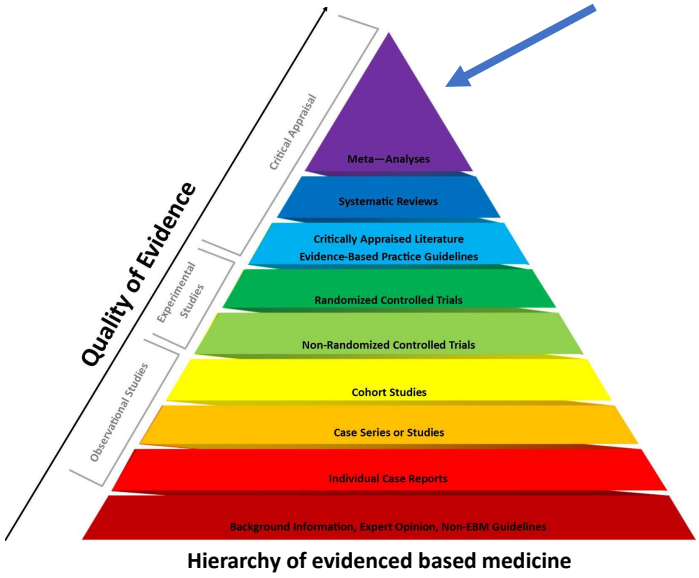
Modern air quality−health effect observational studies perform large numbers of statistical tests using multiple statistical models (MTMM)

1-in-20 results could be '*significant*' (a false positive) even when the null hypothesis is true

Performing many tests on a data set allows researchers to 'select' and 'report' partial results to fit a narrative

- As we will see, a typical air quality−health effect observational study performs large numbers of statistical tests.
- Part of the problem with this is 1-in-20 results could be '*significant*' (a false positive) even when the null hypothesis is true based on statistical theory… *performing many tests can lead to many false positives*.
- Selective reporting of results in a paper – those that are interesting, but false – can fool editors & even peer reviewers of journals.

Meta-analysis

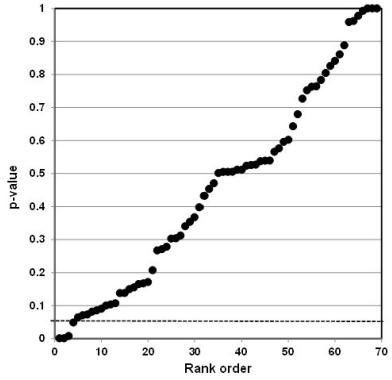


Meta-analysis is a statistical analysis that combines the results of multiple scientific studies:

- It involves searching literature to identify statistically similar studies and then further screening & reviewing to narrow down a list of studies whose data are combined & analyzed (quantitative analysis).
- It is intended to address a single research question… e.g., does *factor A* cause *disease B*?
- Meta-analysis techniques are highly regarded in medical sciences.
- A skilled team of 5−15 researchers can turn out one meta-analysis per week.
- Researchers publish ~5,000 meta-analysis studies per year.
- However, meta-analysis of poorly designed studies produce erroneous statistics and may be misleading.

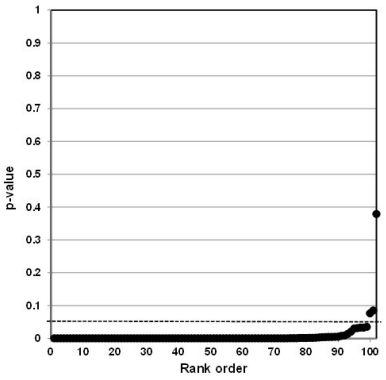
p-value plots – examples of expected behavior

Null hypothesis



(i) elderly long-term exercise training–mortality & morbidity risk (n=69)

Evidence of a real effect



(ii) smoking–squamous cell carcinoma risk (n=102)

What is expected behavior…

- The left plot – a meta-analysis of 69 base studies – forms an approx. 45-degree line providing evidence of randomness—supporting the null hypothesis (no significant association).
- The right plot – a meta-analysis of 102 base studies – forms an approx. line with slope < 1, where most of the p-values are small (< .05), providing evidence for a real effect—supporting a statistically significant association.
- Both plots conform to distinct (single) sample distributions – for null and real effects.
- Large numbers of statistical tests in base studies of a meta-analysis, and whose p-values do not conform to these behaviors should be regarded as suspect.

(1) Orellano et al. 2012 – air quality & mortality

Environment International 142 (2020) 105876

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Review article

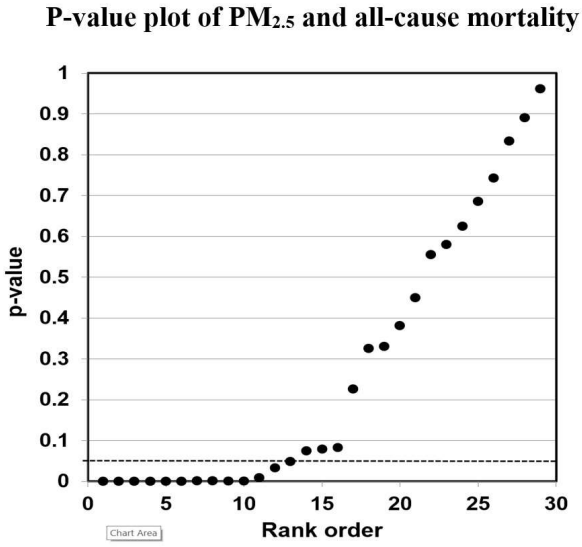
Short-term exposure to particulate matter (PM_{10} and $PM_{2.5}$), nitrogen dioxide (NO_2), and ozone (O_3) and all-cause and cause-specific mortality: Systematic review and *meta*-analysis

Pablo Orellano , Julieta Reynoso , Nancy Quaranta , Ariel Bardach

Claim… "*study found evidence of a positive association between short-term exposure to PM10, PM2.5, NO2, and O3 and all-cause mortality, and between PM10 and PM2.5 and cardiovascular, respiratory and cerebrovascular mortality*"

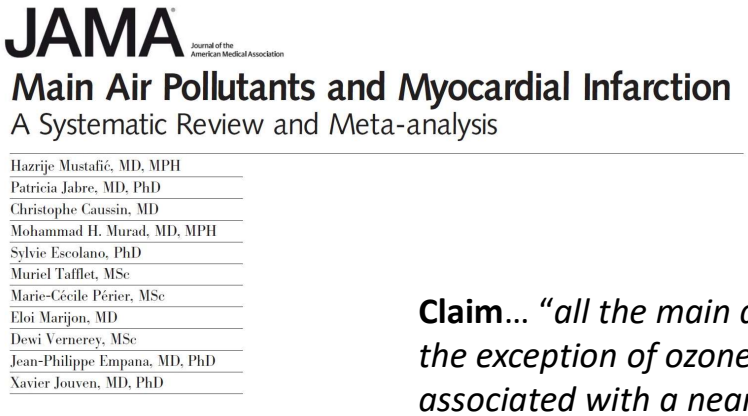
- The Orellano meta-analysis looked at whether 4 outdoor air quality components cause various mortality endpoints.
- Researchers initially identified 2,466 electronic records; they screened 1,632 records.
- They selected 196 studies for quantitative analysis.
- They claimed…

(1) Orellano et al. 2012 – severe testing



- The p-value plot clearly departs from expected behavior – it is bilinear; it breaks into 2 lines.
- This data set is a 2-component distribution.
- As to causes; we see issues such as publication bias, p-hacking, HARKing – well-cited problems in scientific literature – as possible explanations for small p-values

(2) Mustafic et al. 2012 – air quality & heart attack



Claim… "*all the main air pollutants, with the exception of ozone, were significantly associated with a near-term increase in MI* [heart attack] *risk"*

- The Mustafic meta-analysis looked at whether 6 outdoor air quality components cause heart attack (MI).
- Researchers initially identified 1,667 electronic records; they retrieved and reviewed 117 full-text articles.
- They ultimately selected 34 studies for quantitative analysis.
- They stated their study complied with the preferred reporting items of PRISMA; which is an evidence-based minimum set of items for reporting in systematic reviews and meta-analyses.
- They claimed…

(2) Mustafic et al. 2012 – severe testing

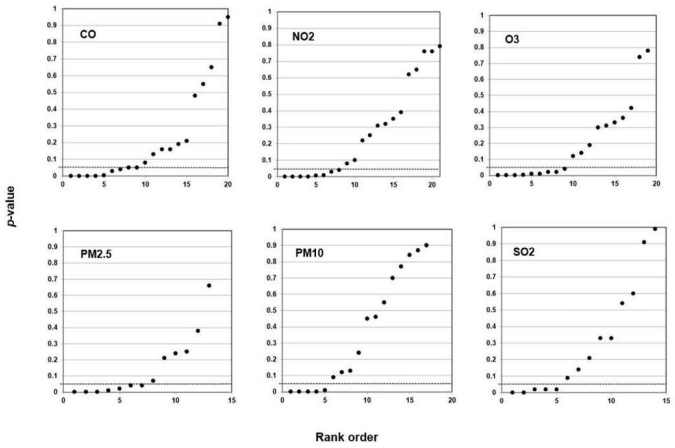
counts

Counts of the number of statistical tests in 34 studies

Statistic	Space1	Space2	Space3
minimum	8	8	240
lower quartile	23	64	2,496
median	36	256	12,288
upper quartile	109	1,024	58,368
maximum	540	16,384	4,587,520

note:
minimum = minimum count
lower quartile = 25th percentile count
median = 50th percentile count
upper quartile = 75th percentile count
maximum = maximum count
Space 1 = Outcomes x Predictors x Lags
Space 2 = $2^{Covariates}$
Space 3 (number of statistical tests) = Space 1 x Space 2

p-value plots



- Counts are summarized on LHS – the median count of statistical tests was over 12,000 from 34 base studies that we reviewed.
- We counted *outcomes*, *predictors*, *time lags* & *covariates* (where appropriate) to estimate numbers of statistical tests.
- The p-value plots – RHS – show no resemblance to expected behavior.
- All of these plots show data sets that are 2-component distributions.
- Their claim should be regarded as suspect particularly with large numbers of statistical tests conducted in their base studies.
- We see similar causes – publication bias, p-hacking, HARKing – as possible explanations for the small p-values

(3) Anderson et al. 2013 – air quality & development of asthma

Air Qual Atmos Health (2013) 6:47–56
DOI 10.1007/s11869-011-0144-5

Long-term exposure to air pollution and the incidence of asthma: meta-analysis of cohort studies

H. Ross Anderson · Graziella Favarato · Richard W. Atkinson

Claim… *"results are consistent with an effect of outdoor air pollution on asthma incidence"*

- The Anderson meta-analysis looked at whether 2 air quality components - PM2.5 & NO2 - associated early childhood lead to asthma later in life.
- They identified 4,165 articles from literature, of which 266 were selected for detailed assessment of full text.
- After further screening for cohorts, they identified 10 articles pertaining to eight "birth cohorts"; and 14 articles pertaining to 9 cohorts with inception in childhood or adult life ("child/adult cohorts").
- They claimed….

(3) Anderson et al. 2013 – severe testing

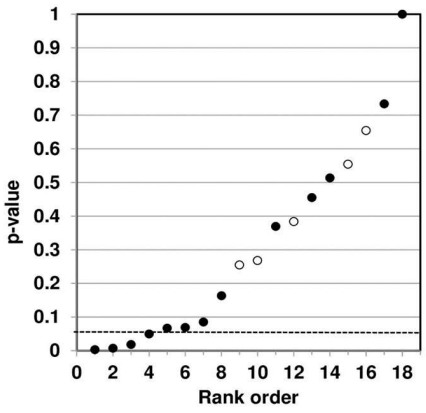
counts

Counts of the number of statistical tests in 19 studies

Statistic	Space1	Space2	Space3
minimum	2	32	96
lower quartile	15	96	1,536
median	24	256	13,824
upper quartile	84	3,072	221,184
maximum	160	262,144	42,000,000

minimum = minimum count
lower quartile = 25th percentile count
median = 50th percentile count
upper quartile = 75th percentile count
maximum = maximum count
Space 1 = Outcomes x Predictors
Space 2 = $2^{Covariates}$
Space 3 (number of statistical tests) = Space 1 x Space 2

p-value plot



Note:
solid circles (●) are NO_2 p-values
open circles (o) are $PM_{2.5}$ p-values

- Counts are summarized on LHS – median count of statistical tests was almost 14,000 from 19 base studies that we reviewed.
- A combined p-value plot shows bilinearily for NO2 (black dots) and ~45 degrees (complete randomness) for PM2.5 (open circles).
- We see a 2-component distribution and little resemblance to expected behavior for NO2.
- Note that we see very good resemblance to expected (null) behavior for PM2.5 – a 'null' effect!

(4) Zheng et al. 2015 – air quality & asthma attack

PLOS ONE

RESEARCH ARTICLE

Association between Air Pollutants and Asthma Emergency Room Visits and Hospital Admissions in Time Series Studies: A Systematic Review and Meta-Analysis

Xue-yan Zheng[1¤a], Hong Ding[1], Li-na Jiang[1], Shao-wei Chen[1], Jin-ping Zheng[2], Min Qiu[1], Ying-xue Zhou[3], Qing Chen[1]*, Wei-jie Guan[2¤b]*

Claim… *"short-term exposure to air pollutants confers an increased risk of asthma-related asthma–related emergency room visits and hospitalizations"*

- The Zheng meta-analysis looked at whether 6 outdoor air quality components cause asthma attack.
- Researchers identified 1,099 literature reports. After screening for titles and abstracts, 246 full-text articles were examined for eligibility
- Ultimately 87 were included for quantitative analysis.
- They stated their study complied with the preferred reporting items of PRISMA.
- They claimed…

(4) Zheng et al. 2015 – severe testing

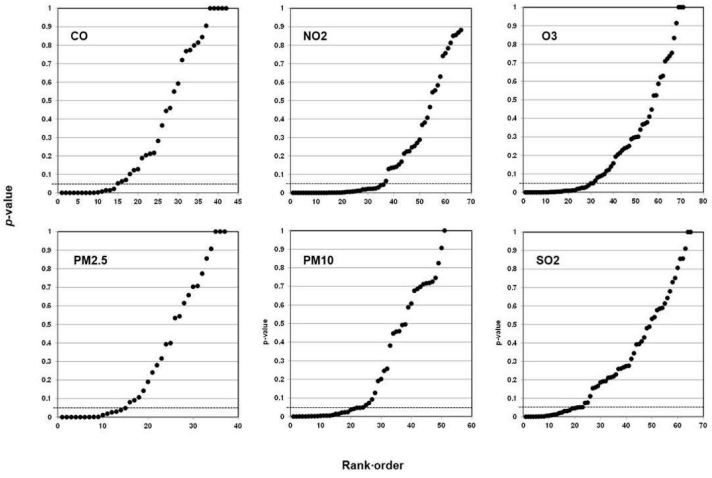
counts

Counts of the number of statistical tests in 17 randomly selected studies from the list of 87

Statistic	Space1	Space2	Space3
minimum	6	4	96
lower quartile	60	16	1,536
median	160	32	15,360
upper quartile	288	256	40,960
maximum	5,120	512	89,600

note:
minimum = minimum count
lower quartile = 25th percentile count
median = 50th percentile count
upper quartile = 75th percentile count
maximum = maximum count
Space 1 = Outcomes x Predictors x Lags
Space 2 = $2^{Covariates}$
Space 3 (number of statistical tests) = Space 1 x Space 2

p-value plots



- Counts on LHS – the median count of statistical tests was over 15,000 from 17 randomly-selected base studies that we reviewed.
- A point to make here… it appears to be typical to perform over 10,000 statistical tests in an environmental epidemiology study.
- p-value plots – look at the lower left plot for PM2.5, we can see many small p-values stacked up below .05 & we also see many p-values exhibiting randomness (>.05)
- All of these plots show data sets that are 2-component distributions.
- Their claim should be regarded as suspect.
- Again, publication bias, p-hacking, HARKing are possible explanations for the small p-values

Publication Bias, p-Hacking, HARKing

Outcome reporting bias: significant results are selected among study outcomes to be published.

Redundant publication bias: significant results are published in more than one paper.

Editors and researchers publish positive finding, $p<0.05$. Anything less typically is not published. There is a bias to positive effects.

If nothing is found, then the data set/question is abandoned.

Bunnies in the sky

"Bunnies in the sky" is a metaphor for a random sighting given many opportunities.

Let's run an epidemiology study!

10-sided dice simulation:

$PM_{2.5}$ causes X

Work Sheet Stan Young, Simulation

MedCondition	YoungFemale	YoungMale	OldFemale	OldMale
1. Angina	.384	.660	.836	.067
2. Arthritis	.180	.251	.098	.451
3. Asthma	.205	.830	.258	.086
4. Cancer	.443	.641	.903	.491
5. C. Bronchitis	.810	.968	.076	.782
6. CHD	.599	.884	.280	.149
7. Emphysema	.100	.861	.107	.999
8. Heart Attack	.797	.543	.622	.158
9. Liver Disease	.183	.334	.596	.466
10. Stroke	.479	.013	.004	.999
11. Thyroid D.	.851	.935	.415	.042
12. Diabetes	.554	.654	.354	.772
13. H. LDL	.537	.383	.475	.900
14. L. HDL	.188	.618	.967	.293
15. C React Protein	.943	.910	.251	.750

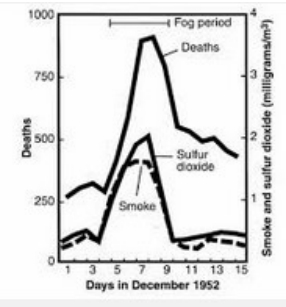
10-sided dice can be used to randomly simulate p-values.
Given are 60 simulated p-values.
We see three below the "magic" 0.05, with one smaller than 0.01.
Three papers could be written.

Smog: Which is different?

Upper left, London Smog 1952. Increase in daily deaths. Upper right, LA 1948. Singapore, Beijing have heavy smog and no report of increased daily deaths.

The etiology

London 1952



Nemery, B., Hoet, P. H. M., Nemmar, A. 2001. The Meuse Valley fog of 1930: an air pollution disaster. *Lancet* 357, 9257: 704–08. https://doi.org/10.1016/s0140-6736(00)04135-0.

Zu, K., Tao, G., Long, C., Goodman, J., Valberg, P. 2016. Long-range fine particulate matter from the 2002 Quebec forest fires and daily mortality in Greater Boston and New York City. *Air Quality, Atmosphere, and Health* 9: 213–21. https://doi.org/10.1007/s11869-015-0332-9.

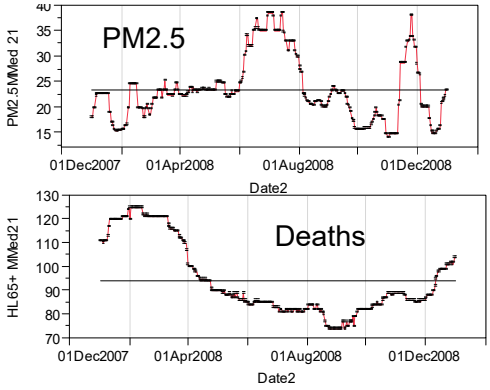
Nemery et al. 2001 noted that there was a combination of factors that lead to deaths in the Meuse Valley. Cold weather. Inversion layer for several days. Acid in the air. Autopsies showed “blown out” lungs. No effect on heart.

Zu et al. 2016 showed PM2.5 increases in Boston and NYC in sync with forest fires in Quebec. There was no increase in daily deaths.

Natural experiments

Wildfires

South Coast, 2008

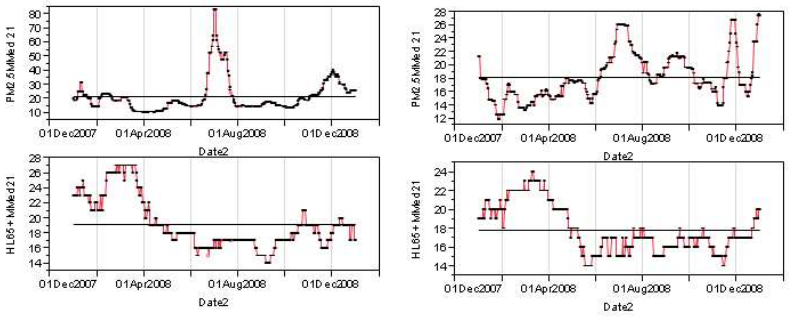


Note the heavy smoke and red fires.
There is an increase in PM2.5.
There is little if any increase in daily deaths.

Natural experiments (2)

Sacramento Valley

San Diego County



Similar lack of effect for all 8 air basins.

Note the spike in PM2.5 around August 1 with much of any effect on daily deaths.

Emotion/Scare

The whole aim of practical politics is to keep the populace alarmed (and hence clamorous to be led to safety) by menacing it with an endless series of hobgoblins, all of them *imaginary*.

H. L. Mencken

Get irate

Flimflam: a deception, a confidence game involving skillful persuasion or clever manipulation of the victim.

21

Humans depend on emotional respond to dangerous situations. But emotion can be exploited. H. L. Mencken: "The whole aim of practical politics is to keep the populace alarmed (and hence clamorous to be led to safety) by menacing it with an endless series of hobgoblins, all of them imaginary."

The medieval church used scare/emotion to control people.

"Global warming/cooling", etc. and "air quality kills" are all likely imaginary.