International Journal of Statistics and Probability; Vol. 11, No. 2; March 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Reliability of a Meta-analysis of Air Quality−Asthma Cohort Studies

S. Stanley Young[1], Kai-Chieh Cheng[2], Jin Hua Chen[2], Shu-Chuan Chen[3] & Warren B. Kindzierski[4]

[1] CGStat, Raleigh, NC, USA

[2] Graduate Institute of Data Science, Taipei Medical University, Taipei City, Taiwan

[3] Department of Mathematics and Statistics, Idaho State University, Pocatello, ID, USA

[4] St Albert, Alberta, Canada

Correspondence: Warren B. Kindzierski, Independent consultant, 12 Hart Place, St Albert, Alberta, T8N 5R1, Canada. Tel: 1-780-458-5921. E-mail: wbk@shaw.ca or warrenk@ualberta.ca

Received: December 23, 2021 Accepted: February 14, 2022 Online Published: February 23, 2022

doi:10.5539/ijsp.v11n2p61 URL: https://doi.org/10.5539/ijsp.v11n2p61

Abstract

What may be a contributing cause of the replication problem in science – multiple testing bias – was examined in this study. Independent analysis was performed on a meta-analysis of cohort studies associating ambient exposure to nitrogen dioxide (NO2) and fine particulate matter (PM2.5) with development of asthma. Statistical tests used in 19 base papers from the meta-analysis were counted. Test statistics and confidence intervals from the base papers used for meta-analysis were converted to p-values. A combined p-value plot for NO2 and PM2.5 was constructed to evaluate the effect heterogeneity of the p-values. Large numbers of statistical tests were estimated in the 19 base papers – median 13,824 (interquartile range 1,536−221,184). Given these numbers, there is little assurance that test statistics used from the base papers for meta-analysis are unbiased. The p-value plot of test statistics showed a two-component mixture. The shape of the p-value plot for NO2 suggests the use of questionable research practices related to small p-values in some of the cohort studies. All p-values for PM2.5 fall on a 45-degree line in the p-value plot indicating randomness. The claim that ambient exposure to NO2 and PM2.5 is associated with development of asthma is not supported by our analysis.

Keywords: cohort studies, air quality, asthma, meta-analysis, multiple testing bias

1. Introduction

1.1 Irreproducible Science

Scholarly publishing in the science, technology and biomedicine fields produced about 2.5 million articles in over 28,000 peer-reviewed journals in 2015 (Ware & Mabe, 2015). Further, Ware & Mabe (2015) indicated continued growth in volumes of these articles at a rate of 3–3.5% per year. Yet research claims in observational studies, randomized trials and, in general, studies across multiple scientific disciplines often do not replicate (Chambers, 2015; Hubbard, 2015; Atmanspacher & Maasen, 2016; NASEM, 2016 & 2019; Harris, 2017; Randall & Welser, 2018; Ritchie, 2020).

The majority of irreproducible studies report positive associations between causative factors (e.g., a behavior or risk factor) and an outcome. Negative (null) studies – those with findings of no associations – are often not reported by researchers (Franco et al., 2014). If negative studies are submitted for publication, editors may reject them out of hand, so a false positive (irreproducible) study can mistakenly be presumed as established fact (Franco et al., 2014; Nissen et al.; 2016).

Published estimates of irreproducible studies or reports range from 51–100% in the biomedical field (Young et al., 2022):

- 41 of 80 studies (51%) examined in the primary care and general medicine field (Glasziou et al., 2008).
- 131 of 257 studies (51%) examined in the clinical psychology, cognitive psychology, cognitive neuroscience, developmental psychology, social psychology, school psychology and various inter-subdisciplinary fields (Hartshorne & Schachner, 2012).
- 129 of 238 studies (54%) examined in the fields of neuroscience, developmental biology, immunology, cell and molecular biology, general biology (Vasilevsky et al., 2013).
- 25 of 45 (56%) clinical studies published in high-impact-factor specialty medical journals in 1990-2003 (Ioannidis,

2005).

- 52 of 63 studies (78%) examined mostly from the oncology field; but several studies were from the fields of women's health and cardiovascular health (Prinz et al., 2011).
- 47 of 53 studies (89%) examined in the fields of haematology, oncology (Begley & Ellis, 2012).
- 52 of 52 studies (100%) examined in the field of nutrition (Young & Karr, 2011).

1.2 Cohort Studies

The beginnings of the cohort studies can be traced to the interest on life behavior and health status information; information that is important to public health (Samet & Munoz, 1988). In a cohort study researchers identify subjects at a point in time when they do not have an outcome of interest (e.g., a disease) and then later to compare the incidence of the disease among groups of exposed and unexposed subjects (Grimes & Schulz, 2002). Periodic follow-up with the subjects can be frequent to record their behaviors and changes in health status (Song & Chung, 2010).

A cohort study can take on a life of its own. The cohort Life Project in England, which examined children born within a narrow period in the 1950s, has become a decades-long study that has provided data for many published articles in a range of social science and health disciplines (Pearson, 2016). Researchers have published 2,500 papers on the 1958 cohort.

1.3 Meta-analysis

A meta-analysis is intended to offer a window into the reliability of a research finding. A meta-analysis examines a research finding by using test statistics from multiple individual studies found in literature (Glass et al., 1981). Two key assumptions of meta-analysis are that the test statistics drawn into the analysis are an unbiased estimate of the effect of interest (Boos & Stefanski, 2013) and that meta-analysis of multiple studies offers a pooled estimate with improved precision (Cleophas & Zwinderman, 2015).

Meta-analyses based on limited evidence, biased studies and/or poor-quality trials are prone to unreliable results (Pereira & Ioannidis, 2011; Packer, 2017). As researchers can often ask a lot of questions and compute many models in an observational study, any statistics coming from such a study may not be unbiased (Young & Kindzierski, 2019). Observational studies that have many hundreds of possible questions at issue may yield extreme findings due to chance. Also, modeling is typically used to reduce variability and aggressive modeling may lead to an underestimate of variability (Schisterman et al., 2009).

Any kind of variability across studies in a meta-analysis may be termed 'heterogeneity' (Higgens & Green, 2011). Heterogeneity occurs because the effects of interest in the subjects studied may not be the same. This can be examined by looking at the 'across study' variability versus the 'within study' variability (Cochran, 1952 & 1954). Very often there is more heterogeneity in meta-analysis than one would expect by chance. One way to deal with heterogeneity is to assume that summary statistics come from a consistent (normal) distribution with extra variability (DerSimonian & Laird, 1986); in which case the meta-analysis process can give a combined (weighted) estimate of an effect. However, the heterogeneity may be more complex and the assumption of selecting values from a normal distribution with extra variability may not be valid for meta-analysis.

1.4 Objective of Study

An independent examination of a meta-analysis drawing statistics from cohort studies was undertaken. The meta-analysis was published by Anderson et al. (2013a,b) and it explored whether associations with ambient air quality early in life lead to development of asthma later in life. As of December 11, 2021, this meta-analysis had 238 *Google Scholar* citations and 135 *Web of Science* citations.

An often-reported cause of the irreproducibility problem is related to researcher statistical methods (Colling & Szucs, 2018). In relation to this, we examined whether asking a lot of questions and computing many models can bias meta-analysis of cohort studies. Asking many questions and computing many models has been referred to as multiple testing and multiple modeling (MTMM) or more generally as multiple testing bias (Westfall & Young, 1993; Young & Karr, 2011). We used analysis search space and p-value plots to independently examine two aspects of the Anderson et al. (2013a,b) meta-analysis:

- Whether research findings in the base papers used for meta-analysis are susceptible to the multiple testing bias.
- Whether heterogeneity in test statistics used for meta-analysis is more complex than simple sampling from a single normal process.

2. Method

Pekkanen & Pearce (1999) note that there are two classes of causes of asthma – primary (related to the increase in risk

of developing the disorder) and secondary (related to asthma attacks or exacerbations). The Anderson et al. (2013a,b) meta-analysis focused on cohort studies of the association between ambient air quality components and development of asthma later in life, and hence on the primary causes of asthma.

A public standard operating procedure (SOP) of the test methods used here was initially filed with the Center for Open Science 'open science framework' (Young, 2019). Anderson et al. conducted a systematic review and meta-analysis of cohort studies of the association between two air quality components – particulate matter with aerodynamic equivalent diameter ⩽2.5 micron (PM2.5) and nitrogen dioxide (NO2) – and incidence of asthma. Incidence was defined as: i) incidence of diagnosed asthma or of new wheeze symptom between two assessments or, ii) in birth cohorts followed up to 10 years of age, a lifetime prevalence estimate of asthma or wheeze symptom.

To increase the number of test statistics for each air quality parameter (PM2.5 & NO2)–outcome pair, Anderson et al. scaled results for studies of particulate matter with aerodynamic diameter <10 μm (PM10) to PM2.5 using a factor of 0.65 and of oxides of nitrogen (NOx) to nitrogen dioxide (NO2) using a factor of 0.44. They indicated that most cohort studies they used for meta-analysis reported test statistics as odds ratios (ORs), but some reported them as relative risks (RRs) or hazard ratios (HRs).

Anderson et al. also indicated that all three quantitative health outcome estimates were combined for their meta-analysis because the outcome of interest (asthma) is quite common, but the effect size is relatively small. A small effect size can be interpreted as a weak relationship between two variables. Their outcomes are referred to here as effect estimates (EEs) with 95% confidence intervals (CIs). These EEs were standardized by Anderson et al. to a 10 $\mu g/m^3$ increment for PM2.5 and NO2.

Anderson et al. identified 17 cohorts in their review. This included eight birth and nine child/adult cohorts of relationships between air quality and incidence of asthma or wheeze symptom with a total of 99 EEs from 24 published studies. Most cohort studies were based on inferred 'within community exposure' contrasts dominated by traffic pollution. Twelve of the 17 cohorts reported at least one positive statistically significant association ($p<.05$) between an air quality component and a measure of asthma incidence. Of the 99 EEs identified, 29 were positive associations and statistically significant (i.e., $p<.05$) and the remaining 70 were null associations.

Thirteen of their cohorts reported results for oxides of nitrogen (NOx), mostly as nitrogen dioxide (NO2), and were used for their meta-analysis of NO2. Of these 13 cohorts, two had multiple publications. Anderson et al. did not state which of the publications they drew upon for their EEs and CIs of the two cohort populations. Also, five cohorts were used for their meta-analysis of PM2.5. Of the five cohorts used, four had multiple publications. Again, Anderson et al. did not state which of the publications they drew upon for their EEs and CIs of the four cohort populations.

It is important to note that epidemiologic studies with null findings more likely remain unpublished compared to studies with positive findings (Chavalarias et al., 2016). Egger et al. (2001) and Sterne et al. (2001) note that this creates a distortion of the literature. This represents a potential problem because a meta-analysis drawing upon test statistics from the literature may only be using misleading, positive findings (Ioannidis, 2008; NASEM, 2019).

Anderson et al. reported the following combined results of their meta-analysis: (i) for the 13 cohort studies with NO2 estimates, the EE was 1.15 (95% CI 1.06 to 1.26) per 10 $\mu g/m^3$, and (ii) for the five cohort studies with estimates for PM2.5, the EE was 1.16 (95% CI 0.98 to 1.37) per 10 $\mu g/m^3$. Finally, Anderson et al. stated in their Abstract… "*The results are consistent with an effect of outdoor air pollution on asthma incidence*."

2.1 Analysis Search Space

Search space counting is introduced as a test of whether studies used in the meta-analysis are susceptible to multiple testing bias. We refer to the cohort studies used for meta-analysis as 'base papers'. Analysis search space (search space counts) represents an estimate of the number of statistical tests performed in a base paper.

Why might this be relevant? There is flexibility available to researchers to undertake a range of statistical tests and use different statistical models in an observational study before selecting, using and reporting only a portion of the test and model results (Young & Kindzierski, 2019). Wicherts et al. (2016) refers to this flexibility as 'researcher degrees of freedom' in the psychological sciences. Base papers with large search space counts suggest the use of a large number of statistical tests and statistical models and the potential for researchers to search through and only report a portion of their results (i.e., positive, statistically significant results).

Analysis search space was estimated for 19 of the Anderson et al. 24 base papers (80%). A listing of the 19 base papers is provided in Appendix A. Electronic copies of these base papers and any corresponding electronic supplementary information files were obtained and read. The number of outcomes, predictors, time lags and covariates reported in each base paper was separately counted as follows (Young & Kindzierski, 2019; Kindzierski et al., 2021):

• The product of outcomes, predictors, and time lags = number of questions at issue (i.e., Questions = outcomes × predictors × lags).

• A covariate may or may not be a confounder to a predictor variable. The only way to test for this is to include/exclude the covariate from a model. As it can be included or excluded, one way to approximate the 'modeling options' is to raise 2 to the power of the number of covariates (i.e., Models = 2^k, where k = number of covariates). Identifying covariates in a published article can be difficult as they might be stated anywhere in the article.

• Questions × Models = an approximation of analysis search space (Search Space).

Three examples of how to estimate analysis search space in observational cohort studies are provided in Appendix B. Estimates of analysis search space are considered to be lower bound approximations (Young & Kindzierski, 2019). What is presented here is based on information that is reported in each base paper evaluated. Finally, we specifically reviewed the 19 base papers focusing our attention on identifying whether a paper: i) discussed/mentioned the multiple testing bias issue in various forms (multiple testing, multiple comparisons, multiplicity) and/or, ii) made any mention of correcting for this issue.

2.2 p-value Plot

Epidemiologic research results have long been required to be statistically significant (NASEM, 1991). Further, environmental epidemiology traditionally uses confidence intervals instead of p-values from a hypothesis test to show statistical significance. As confidence intervals and p-values are derived from the same data set, they are interchangeable, and one can be estimated from the other (Altman & Bland, 2011a,b).

A positive association between two variables in an environmental epidemiology study can be considered statistically significant where the confidence interval for a test statistic excludes the null hypothesis or the p-value is less than .05. The p-value is a random variable derived from a distribution of the test statistic used to analyze data and to test a null hypothesis (Kindzierski et al., 2021). The p-value can be defined as the probability, if nothing is going on, of obtaining a result equal to or more extreme than what was observed.

Hung et al. (1997) indicate that under the null hypothesis, the p-value is distributed uniformly over the interval 0 to 1 regardless of sample size. A distribution of true null hypothesis points in a p-value plot should form a straight line (Schweder & Spjøtvoll, 1982). A plot of rank-ordered p-values related to true null hypothesis points should conform to a near 45-degree line (Westfall & Young, 1993). The plot can be used to assess the validity of a false finding being taken as true and can be used to test the reliability of the findings made in base papers used for meta-analysis.

A p-value plot was constructed using 18 Anderson et al. (2013a,b) p-value estimates – 13 for NO2 and five for PM2.5 – and interpreted after Schweder & Spjøtvoll (1982) and Young & Kindzierski (2019):

• The p-values were computed from the EEs and CIs assuming symmetrical CIs using JMP statistical software (SAS Institute, Cary, NC).

• The p-values were ordered from smallest to largest and plotted against the integers, 1, 2, 3, …

• If p-value results are random (i.e., a true null relationship), the p-value plot should roughly follow a 45-degree line indicating a uniform distribution.

• Alternatively, p-values should be on a roughly straight line with a slope considerably less than 45 degrees if a true relationship exists.

• If analysis search space counts are high and the corresponding plotted p-values exhibit a two-component – bilinear shape – then the p-values used for meta-analysis comprise a mixture and a general (over-all) finding is not supported. In addition, the p-value reported for the combined statistic of the meta-analysis is not valid. This is elaborated further in the study.

To assist in interpretation of the visual behavior of p-value plots, plots for 'plausible true null' and 'plausible true alternative' hypothesis outcomes based on meta-analysis of observational datasets were constructed (Appendix C). Hung et al. (1997) note the distribution of the p-value under the alternative hypothesis – where p-values are a measure of evidence against the null hypothesis – is a function of both sample size and the true value or range of true values of the tested parameter. The p-value plots presented in Appendix C represent examples of distinct (single) sample distributions for each condition – i.e., for true null associations and true effects between two variables. Evidence for p-value plots exhibiting behaviors outside of that shown in Appendix C should be treated as questionable particularly where analysis search space counts are high.

3. Results

3.1 Analysis Search Space

Estimated analysis search spaces for 19 base papers we examined from Anderson et al. (2013a,b) are presented in Table 1. These 19 papers represented 14 of the 17 cohort studies used by Anderson et al. for their meta-analysis. From Table 1, investigating multiple – 2 or more – asthma outcomes (i.e., Outcomes) in the cohort studies were as common as single outcome investigations. In addition, use of multiple Predictors and Lags was common. So was adjusting for multiple possible Covariate confounders. Examining multiple factors (i.e., outcomes, predictors, lags and covariates) seemingly represents a reasonable attempt to simulate/model possible exposure−disease combinations, however these multiple combinations can inflate the overall number of statistical tests performed in a single study.

Table 1. Counts and analysis search spaces for 19 base papers considered by Anderson et al. (2013a,b) in their meta-analysis

RowID	Study cohort	Outcomes	Predictors	Lags	Covariates	Questions	Models	Search space
1	BAMSE	7	3	4	6	84	64	5,376
2	British Columbia	1	8	4	7	32	128	4,096
3	CHS	1	2	8	15	16	32,768	524,288
4	CHS	1	6	5	10	30	1,024	30,720
5	CHS 2003	1	5	3	15	15	32,768	491,520
6	CHIBA	3	1	3	6	9	64	576
7	CHIBA	1	3	6	6	18	64	1,152
8	CHIBA	5	4	4	8	80	256	20,480
9	ECHRS	1	1	6	11	6	2,048	12,288
10	GINIplus+LISAplus	4	4	6	12	96	4,096	393,216
11	MISSEB	1	2	7	6	14	64	896
12	OLIN	1	3	4	5	3	32	96
13	OSLO	4	2	3	11	24	2,048	49,152
14	PIAMA	8	4	4	18	128	262,144	33,000,000
15	PIAMA	5	4	8	18	160	262,144	42,000,000
16	RHINE	1	2	1	8	2	256	512
17	TRAPCA	6	3	6	7	108	128	13,824
18	TRAPCA	7	3	4	9	84	512	43,008
19	AHSMOG	1	3	3	7	15	128	1,920

Notes. Refer to Appendix A for a listing of the 19 study cohort names; Questions = Outcomes × Predictors × Lags; Models = 2^k where k = number of Covariates; Search space = approximation of analysis search space = Questions × Models; none of the papers made any adjustments/corrections for multiple testing bias.

None the 19 base papers we reviewed made any adjustments/corrections for multiple testing bias in their analysis. Seventeen of the 19 papers made no mention of the issue. One paper (Morgenstern et al., 2007) stated they… *did not adjust for multiple testing* in their discussion. Another paper (McDonnell et al., 1999) used Bonferroni's correction but only to explain similarities in characteristics of four different groups making up their study cohort. They did not make any adjustment/correction for multiple testing bias in their subsequent analysis.

Summary statistics of possible numbers of statistical tests performed in the 19 base papers are presented in Table 2. The median number (interquartile range, IQR) of Questions and Models was 24 (IQR 15−84) and 256 (IQR 96–3,072), respectively. The median number (IQR) of possible statistical tests (Search space) of the 19 base papers was 13,824 (IQR 1,536−221,184).

Given the large numbers of possible tests, the statistics drawn from the cohort studies are unlikely to offer unbiased statistics for meta-analysis. Although not shown, covariates in each of the cohort studies vary considerably from study to study (the reader is referred to the original Anderson et al. 2013a supplemental files). For comparison purposes,

search space counts of air quality component−heart attack observational studies are also large – i.e., median (IQR) = 6,784 (2,600−94,208), n=14 (Young et al., 2019), and = 12,288 (2,496−58,368), n=34 (Young & Kindzierski, 2019).

Table 2. Summary statistics for counts estimated for 19 base papers considered by Anderson et al. (2013a,b) in their meta-analysis

Statistic	Outcomes	Predictors	Covariates	Lags	Questions	Models	Search space
Minimum	1	1	1	5	2	32	96
Lower quartile	1	2	4	7	15	96	1,536
Median	1	3	4	8	24	256	13,824
Upper quartile	5	4	6	12	84	3,072	221,184
Maximum	8	8	8	18	160	262,144	42,000,000

Notes. Questions = Outcomes × Predictors × Lags; Models = 2^k where k = number of Covariates; Search space = approximation of analysis search space = Questions × Models.

3.2 p-value Plot

Table 3 presents EEs, CIs and calculated p-values for 18 cohort studies used in their meta-analysis calculation. A plot of the sorted p-values versus the integers is given in Figure 1. Both p-values for NO2 (indicated by solid circles, ●) and PM2.5 (indicted by open circles, o) are combined in Figure 1. This (combining results for NO2 and PM2.5) is the same as what Anderson et al. (2013a) did to compute their meta-analytic statistic.

Table 3. Effect estimate (EE), lower confidence level (CL_{low}) and upper confidence level (CL_{high}) values and corresponding p-values estimated for cohort studies used by Anderson et al. (2013a,b) in their meta-analysis

Air Component	Study cohort, outcome	EE	CL_{low}	CL_{high}	p-value
NO2	BAMSE, wheeze	1.01	0.98	1.04	0.5135
	British Columbia, asthma	1.13	1.04	1.23	0.0073
	CHS 2003, asthma	1.03	1.01	1.05	0.0033
	CHS, asthma	1.24	1.06	1.46	0.0187
	CHIBA, asthma	1.32	1.02	1.71	0.0691
	ECRHS, asthma	1.43	1.02	2.00	0.0854
	KRAMER, asthma	1.19	0.85	1.68	0.3695
	MISSEB, asthma	1.32	0.73	2.41	0.4553
	OLIN, asthma	1.00	0.35	2.87	1.0000
	Oslo Birth Cohort, asthma	0.93	0.85	1.00	0.0673
	PIAMA, asthma	1.16	0.96	1.41	0.1634
	RHINE, asthma	1.46	1.07	1.99	0.0500
	TRACPA, asthma	0.71	0.14	3.48	0.7336
PM2.5	AHSMOG, asthma	1.08	0.85	1.38	0.5541
	British Columbia, asthma	1.10	0.90	1.35	0.3837
	CHIBA, asthma	1.86	0.90	3.86	0.2547
	PIAMA, asthma	2.06	0.91	4.66	0.2678
	TRACAP, asthma	1.60	0.45	5.70	0.6542

Notes. Study cohorts presented here are in the same order as presented in the Anderson et al. (2013a) Fig. 1 for NO2 study cohorts and Fig. 2 for PM2.5 study cohorts.

The Figure 1 relationship presents as bi-linear – six p-values are near or below nominal significance (.05) and the remaining p-values >.05 fall on an approximately 45-degree line. This is different from behavior of both plausible true null and true alternative hypothesis outcomes (Appendix C Figs. C−1 and C−2). A plausible true null hypothesis outcome presents as a sloped line from left to right at approximately 45 degrees, and a plausible true alternative hypothesis outcome presents as line with a majority of p-values below .05 in p-value plots.

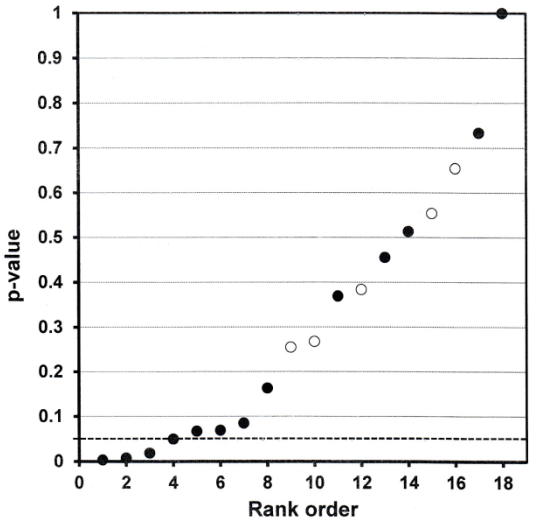

Figure 1. P-value plot for the Anderson et al. (2013a,b) meta-analysis (note: solid circles (●) are NOx p-values; open circles (o) are PMx p-values)

The p-value plots are basic technology used by others (e.g., Selwyn, 1989; Westfall & Young, 1993; Cao et al., 2007; Young et al. 2009; Ryan et al., 2013; Young & Kindzierski, 2019; Kindzierski et al., 2021). The two-component mixture of p-values in Figure 1 may be a combination of studies showing an association and no association, but both outcomes cannot be true. Questionable research practices (QRP) involve approaches used by researchers during data collection, analysis and reporting that can increase false-positive findings in published literature (de Vrieze, 2021; Ravn & Sørensen, 2021). QRP cannot be ruled out as an explanation for small p-values in several of the cohort studies in Table 3 and Figure 1. A p-value (mixture) relationship like this does not support a finding that exposure to ambient NOx and PM2.5 early in life is associated with development of asthma later in life.

Higgens and Green (2011) assert that heterogeneity will always exist in meta-analysis whether or not one can detect it using a statistical test. Statistical heterogeneity (I^2) quantifies the proportion of the variation in point estimates due to among−study differences. I^2 is a standard measure for heterogeneity and Anderson et al. (2013a,b) reported an I^2=64.1% (p<.001) for NO2 based on 13 study cohorts and I^2=7.4% (p=.364) for PM2.5 based on 5 study cohorts.

4. Discussion

Two statistical methods – analysis search space and p-value plots – were used to independently test the reliability of a meta-analysis of cohort studies published by Anderson et al. (2013a,b). Large search space counts without evidence of corrections for multiple testing bias in base papers is one measure for identifying limitations of a meta-analysis. Estimated analysis search spaces of the base studies used by Anderson et al. (2013a,b) – Tables 1 and 2 – indicates that there were large numbers of possible statistical tests performed in the base studies. Results taken from these studies are unlikely to offer unbiased measures for meta-analysis.

The p-value plot constructed for statistics taken from the base studies (Figure 1) shows a two-component mixture. This bi-linear pattern/shape is clearly different with p-value plot behavior of either plausible true null and true alternative hypothesis outcomes (Figs. C−1 and C−2). Taken together, this evidence does not support the meta-analysis as being a reliable study for other researchers to depend upon. This is discussed in more detail.

4.1 Interpretation of Anderson et al. Meta-analysis

The overall approach and EEs and CIs in the Anderson et al. (2013a,b) meta-analysis is taken at face value and interpretations of our independent tests are made from there. As to an air quality−asthma development relationship, there may also be other possible explanations for statistical associations in observational studies. Some of these explanations relate to study methodology and include (Clyde, 2000; Pocock et al., 2004; Ioannidis, 2008; Sarewitz, 2012; Chambers, 2015; Simonsohn et al., 2014; Hubbard, 2015; Harris, 2017; Young & Kindzierski 2019):

- Improper selection of datasets for analysis.
- Improper selection of statistical models.
- Flexible choices in methods to compute statistical results, including undertaking multiple testing and multiple modeling without statistical correction.

- Inadequate treatment of confounders and other latent variables.
- Selective reporting of results.
- Publication bias, non-reporting of null results.

There are many aspects of choice involved in modeling air quality−health effect relationships in observational studies (Young & Kindzierski, 2019). Some of these choices involve which parameters and confounding variables (covariates) to include in a model, what type of lag structure for covariates to use, which interactions need to be considered, and how to model nonlinear trends (Clyde, 2000). Because of the many potential parameters and confounders that may be included in a study, some aspect of model selection is often used. Even if models are selected in an unbiased manner, different model selection strategies may lead to very different models and outcomes for the same set of data. On the other hand, inherent bias may lead researchers to choose models that provide selective outcomes (Young & Kindzierski, 2019).

The p-value plot test and the resulting mixture relationship of p-values from the cohort studies (Figure 1) does not support a general air quality−asthma incidence relationship. Although Anderson et al. follow a typical statistical approach for meta-analysis, their approach will not be meaningful if EEs & CIs drawn from base studies are not unbiased and/or if the test statistics drawn from the base studies as a whole form a two-component mixture.

We consider the Anderson et al. study as a standard meta-analysis… a research question is selected, a computer search is undertaken for relevant published papers involving study cohorts, papers are identified, filtered, and a final set of base papers is selected. The etiology they examined is… *whether ambient air quality early in life leads to development of asthma later in life*. Each cohort study (i.e., base paper) they selected looked at air quality, including one or more of the following air components – carbon monoxide, nitrogen dioxide, sulfur dioxide, ozone, particulate matter (PM).

However, each cohort study varied in terms of the specific air components studied. They identified 18 study cohorts with a total of 99 EEs that examined air quality and asthma, but they only ended up doing a formal meta-analysis on NOx (NO or NO2) and PMx (PM10 and/or PM2.5). Three outcome estimates (ORs, RRs or HRs) with upper and lower confidence limits were extracted from the base papers and a random effects analysis assuming the statistics were normally distributed was used after DerSimonian & Laird (1986).

The Anderson et al. initial computer search identified 4,165 possibly relevant papers. From this, 266 papers were examined in detail and 13 cohorts studies were selected that reported on NOx and five cohort studies were selected that report on PM. A numerical meta-analysis was computed on the two datasets, NOx and PMx, separately and computed p-values for these datasets and combined the p-values into one figure (Figure 1).

In their search, Anderson et al. also identified asthma-related effect studies for other air quality components – e.g., carbon monoxide, ozone and sulfur dioxide. They did not explain reasons for excluding these components in the meta-analysis. As for NO2 and PM, the search space counts – numbers of possible statistical tests conducted – in the selected base studies (Tables 1 and 2) are considered large – median search space 13,824 (IQR 1,536−221,184). Given such large search spaces, there is little assurance that test statistics drawn from the base papers into their meta-analysis are unbiased.

4.2 Heterogeneity

Regarding the interpretation of quantitative measures of heterogeneity (I^2), Higgens et al. (2003) assign low, moderate and high I^2 values of 25%, 50%, and 75% for meta-analysis. Higgens and Green (2011) provide other guidance for interpretation: 0−40% may not be important, 30−60% may represent moderate heterogeneity, 50−90% may represent substantial heterogeneity and 75−100% represents considerable heterogeneity. The Higgens and Green (2011) and Higgens et al. (2003) guidance suggests that the Anderson et al. (2013a,b) meta-analysis of 13 NO2 study cohorts is associated with moderate to substantial heterogeneity.

Forstmeier et al. (2017) note that a key source of heterogeneity in meta-analysis is publication bias favoring positive effects, often due to researcher degrees of freedom (flexibility) to find a statistically significance effect more often than expected by chance As for the Anderson et al. meta-analysis, it is noted that heterogeneity in their NO2 dataset is not simply due to an increase in across-study (study-to-study) variability – but is a much more problematic two-component mixture (see Figure 1). Specifically, some NO2 studies have very small p-values that may suggest real causal relationships or QRP, whereas other p-values fall on a 45-degree line indicating randomness (i.e., no relationship at all). As for their meta-analysis of 5 PM2.5 study cohorts, all corresponding p-values fall on a 45-degree line indicating complete randomness (Figure 1).

There are two standard approaches to heterogeneity: 1) Find the covariates that give rise to the heterogeneity. Covariates are typically examined and usually they do not remove the heterogeneity. 2) Use the random effects analysis model of

DerSimonian & Laird (1986). This approach assumes that there is one normal distribution and a contribution of study-to-study variability, and that bias is not a big issue. This approach was specifically developed for randomized control trials (RCTs) where these assumptions are more reasonable. One normal distribution means one etiology with superimposed study-to-study variability.

However, the Anderson et al. meta-analysis is not of RCTs; it combines observational studies, and a mixture is observed (Figure 1) – some studies positive and some studies null. There is a difference between sampling from a normal distribution and a two-component mixture. Publication bias and multiple testing bias may, in part, explain the nature of the heterogeneity – i.e., two-component mixture (Figure 1) – for NO2. These factors cannot be dismissed as possible explanations for their findings.

A further possible hidden problem is that meta-analysis assumes that heterogeneity among statistics from relevant base papers is randomly distributed around the true value (Charlton, 1996). This holds true if errors across base papers used for meta-analysis are balanced – i.e., errors in some base studies in one direction are cancelled or balanced out by errors in other base studies in the other direction. Therefore, statistical pooling and averaging in meta-analysis in theory produces an error-reduced estimate of the underlying, unbiased, 'true' value.

However, pooling statistics from base papers in meta-analysis unavoidably includes hidden biases of the individual studies. Given pervasiveness of bias in science today (Sarewitz, 2012; Forstmeier et al., 2017), a more likely situation is that most researchers tend to make the same errors in the same direction – i.e., their test statistics have similar biases related to seeking out positive associations (which is more publishable). Such a condition violates a key assumption of meta-analysis and any statistic under this situation is not meaningful.

4.3 Real Versus Random (Chance) Associations

The p-value plot (Figure 1) exhibits a bi-linear appearance and is clearly different from p-value plots of both plausible true null and true alternative hypothesis outcomes (Appendix C Figs. C−1 and C−2). Six p-values are below or near .05, a value often taken as ‘statistically significant’, and twelve p-values appear completely random – a two-component mixture. Firstly, any sort of statistical averaging – weighted or not – for a mixture of this type in inappropriate. Secondly, both findings cannot be true. Evidence for real versus random statistical associations is considered further here (refer to Table 4). The factors presented in Table 4 are intended as a checklist related to methodology of the base papers for researchers to assist in the interpretation of real versus random statistical associations.

Table 4. Factors to consider when evaluating meta-analysis results presenting as bi-linear in a p-value plot

Possibility 1	**Possibility 2**
Small p-values true	**Small p-values false**
Most of the base papers show a small p-value	Evidence of QRP in base papers
There will be supporting literature	Covariates correlated with outcome, bias
There will be a reasonable etiology	Very large sample size elevates small bias to cause
No evidence of QRP in base papers	A number of Bradford Hill criterion not met
Most Bradford Hill criterion met	
Large p-values false	**Large p-values true**
Poor research technique in base papers	Distribution of large p-values is uniform
Underpowered studies	Good negative effect studies
Masking covariates hide real effect	No clear etiology
Role of chance	

Notes. QRP = questionable research practices.

4.3.1 Small p-values True & Large p-values False

First, suppose the small p-values represent true associations, i.e., there is a real air quality−asthma association. In this study, there are two small p-values – .00328 and .00732 (Table 3). P-values this small are often interpreted by researchers as being real. These two p-values are close to a .005−action level proposed by Johnson (2013), but larger than a .001−action level proposed by Boos and Stefanski (2011) for making such an interpretation. Here, the term ‘action level’ means that if the study is replicated, the replication will give a p-value less than .05.

These rules of thumb – the traditional .05 and more-recently proposed .005 and .001 p-value decision criteria – all presume only one statistical test and one p-value result, i.e., no multiple testing bias issues, and that the result is from a well-conducted study (i.e., with randomization, blinding and blocking). This is not true for the Anderson et al. (2013a,b) meta-analysis and its base studies. The median number of possible p-values over the base studies is 13,824. A small p-value can easily arise by chance given this many tests (Bock, 2016). When researchers have many different hypotheses and carry out many statistical tests on the same set of data, they run the risk of concluding that there are real differences or real associations when in fact there are none (Kavvoura et al., 2007).

Yet there is an abundance of published literature suggesting an overall statistical air quality component−adverse health relationship. There are large numbers of papers reporting positive statistical associations between some air quality variable and a health effect. For example, a Google Scholar search of the exact phrase "air pollution" and term "asthma" in a title over the years 1990−2020 returned 1,330 hits (search done 14 December 2021).

If small p-values are true, plausible explanations are needed for large p-values in Figure 1 being false. Here, one can speculate it is possible that some of the papers have a large p-value due to poor data, methods, small sample size or just chance. However, seven of 13 NO2 p-values and all five PM2.5 p-values are greater than .05 (Table 3 and Figure 1). This requires further rationalization given the presumed careful procedure used by Anderson et al. researchers to screen and select their study cohorts and base papers for meta-analysis.

4.3.2 Small p-values False & Large p-values True

On the other side of the coin is a possibility that the small p-values are false. How might this be the case? In the presence of large numbers of statistical tests performed in the base studies, two plausible ways related to methodology are offered which may contribute to a meta-analysis failing:

• P-hacking in base studies (Streiner, 2018). P-hacking is multiple testing and multiple modeling without statistical correction (Chambers, 2015; Hubbard, 2015; Harris, 2017). Search space counts of base studies aids understanding of this issue. Specifically, p-hacking cannot be ruled out as an explanation for small p-values coming from base studies where no statistical corrections are made for large numbers of test performed.

• Not properly controlling for covariates in base studies (Brenner, 1998; Wang & Yin 2013) such that controlling for them may make the small p-values disappear.

While both are important, p-hacking may be serious in published biomedical studies. For example, Hayat et al. (2017) randomly sampled and reviewed 216 of 1,023 published articles from seven top tier general public health journals for the year 2014 with an objective quantifying basic and advanced statistical methods used in public health research. These journals included: *Epidemiology*, *American Journal of Epidemiology*, *American Journal of Public Health*, *Bulletin of World Health Organization*, *European Journal of Epidemiology*, *American Journal of Preventive Medicine* and *International Journal of Epidemiology*. They reported that statistical corrections for multiple testing bias were only made in 5.1% of the 216 studies they reviewed (i.e., ~1−in−20 published studies).

We reviewed the Hayat et al. (2017) Supplemental Information and we emailed the corresponding author – Hayat – in attempts to identify which articles indicated adjustments for multiple testing bias. Both the Supplemental Information and email response provided by Hayat indicated 10 (not 11) or 4.6% of the 216 randomly sampled articles made adjustments for multiple testing bias. It is further speculated by us that the articles making adjustments may be genetic rather than traditional epidemiology articles, and that published traditional epidemiology articles making adjustments for multiple testing bias may be much less than 4.6%.

Kavvoura et al (2007) noted that there is an apparent tendency among epidemiology researchers to avoid making statistical corrections for multiple testing bias, highlighting statistically significant findings, and avoiding highlighting nonsignificant findings in their research papers. This behavior may be a problem, because many of these significant findings could in future turn out to be false positives.

4.4 Testing of Meta-analysis Claims

It has been stated previously that the body of literature available for meta-analysis may be distorted with positive−association studies and a systematic review or meta-analysis of these studies may be biased (Egger et al., 2001; Sterne et al., 2001) because researchers are summarizing information and data from a misleading, selected body of evidence (Ioannidis, 2008; NASEM, 2019). We believe this applies to the epidemiolocal literature, and this alone supports a need for testing of claims made in meta-analysis of this literature. Mayo (2018) endorses Karl Popper's approach of developing scientific knowledge by identifying and correcting errors through strong (severe) tests of scientific claims. We also support this approach.

It makes sense to step back from a detailed consideration of the air quality component−asthma claim made by Anderson

et al. (2013a,b). A scientist is expected to make a good case for a research claim, and it ought to survive a battery of severe but passable tests. Several examples exist in epidemiolocal literature where air quality component−chronic disease claims – both positive and null associations – have been independently tested.

While there is observational evidence that long-term exposure to particulate matter (PM) is associated with premature death in urban populations, confounding by unmeasured variables remains a valid concern in observational studies. Greven et al. (2011) attempted to independently replicate a long-term particulate matter exposure−acute mortality claim using the US Medicare Cohort Air Pollution Study (MCAPS) dataset. This dataset included individual−level information on time of death and age on a population of 18.2 million for the period 2000–2006.

Greven et al. suggested two ways to make a good case for a positive air quality component−health effect claim is to test for 'within location' or 'across location' effects. Positive 'across location' effects might be due to confounding whereas positive 'within location' effects would be less likely biased by confounding. Using this within location approach, Greven et al. was unable to replicate the long-term particulate matter exposure−acute mortality claim.

Another example is with the Young et al. (2017) analysis of an air quality−acute death claim for the eight most populous California air basins. This analysis included over 2,000,000 deaths and over 37,000 exposure days over a 13-year period. Young et al. examined each air basin individually (i.e., 'within location' analysis) and observed null effects like Greven et al (2011). Here one must keep in mind that as sample size goes to infinity, the standard error (SE) goes to zero. So, any small but statistically significant 'across location' effect observed between two air basin populations, each with large sample sizes, has a good chance of being due to bias.

This bias can largely be controlled using a method referred to as Local Control (Obenchain & Young, 2017). One clusters objects into many small clusters and does an analysis within each cluster. One can then observe how the analysis result changes (or does not change) across clusters. Obenchain & Young applied Local Control to a historical air quality (total suspended particulate−mortality) dataset describing a 'natural experiment' initiated by the federal Clean Air Act Amendments of 1970 (specifically, the Chay et al., 2003 dataset).

Chay et al. (2003) used a comprehensive county-level (US) dataset available compiled on population, mortality, total suspended particulate matter (TSP) levels and economic conditions for the period 1969–1974. Obenchain & Young replicated the Chay et al. finding of a no TSP−mortality association. Thus, the control of confounding is important. Variables can be put into a model or an analysis and be restricted to limited geographic regions (e.g., clusters) thereby reducing the influence of confounding factors.

Two statistical methods are demonstrated here as a form of testing of scientific claims made in meta-analysis of observational cohort studies:

- Search space counting and identifying whether corrections for multiple testing bias are made in base papers is one measure. Search space counting allows one to obtain a clearer picture of the numbers of statistical tests that may have been performed in base studies. Test statistics drawn from base studies with large search space counts are unlikely to offer unbiased measures for meta-analysis where no corrections are made for the number of statistical tests performed.
- Examining the behavior of p-values in a ranked plot (p-value plot) for test statistics drawn from base studies into a meta-analysis is another measure. P-value plotting provides a test of results where the underlying data itself remains hidden.

These tests enable a user to independently diagnose specific meta-analysis claims to judge the potential for use of QRP such as multiple testing bias, p-hacking, publication bias (Banks et al., 2016). The Anderson et al. (2013a,b) meta-analyses associating ambient air quality early in life with development of asthma later in life failed these tests. The p-value plot constructed for statistics used from the base studies by Anderson et al. (2013a,b) showed a two-component mixture. This bi-linear pattern/shape is clearly different with p-value plot behavior of both plausible true null and plausible true alternative hypothesis outcomes.

5. Findings

Estimation of analysis search spaces of 19 base papers used in the Anderson et al. meta-analysis indicated that the numbers of statistical tests possible were large – median 13,824 (interquartile range 1,536−221,184; range 96−42M) in comparison to actual statistical test results presented. Given such large search spaces, there is little assurance that test statistics drawn from the base papers into the meta-analysis are unbiased.

A p-value plot showed that heterogeneity of the NO2 results across studies is consistent with a two-component mixture. Meta-analytic averaging across a mixture is inappropriate. The shape of the p-value plot for NO2 appears consistent with use of questionable research practices to obtain small p-values in several of the cohort studies. As for PM2.5 results, all corresponding p-values fall on a 45-degree line in the p-value plot indicating complete randomness rather

than a true association.

Anderson et al. claim an association of air quality and development of asthma. Our analysis does not support their claim. Because of multiple testing bias, it cannot be ruled out that test statistics drawn from the base papers and used for meta-analysis by Anderson et al. are unbiased. Also, heterogeneity of test statistics across base papers used for the meta-analysis is more complex than simple sampling from a normal process.

Acknowledgments

No external funding was provided for this study. The study was conceived based on previous work undertaken by CG Stat for the National Association of Scholars (nas.org), New York, NY.

References

Altman, D. G., & Bland, J. M. (2011a). How to obtain a confidence interval from a P value. *British Medical Journal, 343*, d2090. https://doi.org/10.1136/bmj.d2090

Altman, D. G., & Bland, J. M. (2011b). How to obtain the P value from a confidence interval. *British Medical Journal, 343*, d2304. https://doi.org/10.1136/bmj.d2304

Anderson, H. R., Favarato, G., & Atkinson, R. W. (2013a). Long-term exposure to air pollution and the incidence of asthma: meta-analysis of cohort studies. *Air Quality, Atmosphere, and Health, 6*, 47–56. https://doi.org/10.1007/s11869-011-0144-5

Anderson, H. R., Favarato, G., & Atkinson, R. W. (2013b). Erratum to: Long-term exposure to air pollution and the incidence of asthma: meta-analysis of cohort studies. *Air Quality, Atmosphere, and Health*, *6*, 541–542. https://doi.org/10.1007/s11869-012-0184-5

Atmanspacher, H., & Maasen, S. (eds). (2016). *Reproducibility: Principles, problems, practices, and prospects*. Hoboken, NJ: John Wiley & Sons.

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Editorial: Evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business and Psychology*, *31*(3), 323–338. https://doi.org/10.1007/s10869-016-9456-7

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature, 483*, 531–533. https://doi.org/10.1038/483531a

Bock, E. (2016). Much biomedical research is wasted, argues Bracken. *NIH Record*, July 1, 2016 Vol. LXVIII, No. 14. https://nihrecord.nih.gov/2016/07/01/much-biomedical-research-wasted-argues-bracken

Boos, D. D., & Stefanski, L. A. (2011). P-value precision and reproducibility. *The American Statistician, 65*(4), 213−221. https://doi.org/10.1198/tas.2011.10129

Boos, D. D., & Stefanski, L. A. (2013). *Essential statistical inference: Theory and methods*. New York, NY: Springer.

Brenner, H. (1998). A potential pitfall in control of covariates in epidemiologic studies. *Epidemiology*, *9*(1), 68–71. https://doi.org/10.1097/00001648-199801000-00014

Cao, H., Hripcsak, G., & Markatou, M. (2007). A statistical methodology for analyzing co-occurrence data from a large sample. *Journal of Biomedical Informatics, 40*, 343–352. https://doi.org/10.1016/j.jbi.2006.11.003

Chambers, C. (2015) *The seven deadly sins of psychology, A manifesto for reforming the culture of scientific practice*. Princeton, NJ: Princeton University Press.

Charlton, B. C. (1996). The uses and abuses of meta-analysis. *Family Practice*, *13*(4), 397–401 https://doi.org/10.1093/fampra/13.4.397

Chavalarias, D., Wallach, J. D., Li, A. H., & Ioannidis, J. P. (2016). Evolution of reporting p values in the biomedical literature, 1990−2015. *Journal of the American Medical Association, 315*(11), 1141−1148. https://doi.org/10.1001/jama.2016.1952

Chay, K., Dobkin, C., & Greenstone, M. (2003). The Clean Air Act of 1970 and adult mortality. *Journal of Risk & Uncertainty*, *27*, 279–300. https://doi.org/10.1023/A:1025897327639

Cleophas, T. J., & Zwinderman, A. H. (2015). *Modern meta-analysis: Review and update of methodologies*. New York, NY: Springer.

Clyde, M. (2000). Model uncertainty and health effect studies for particulate matter. *Environmetrics*, *11*, 745−763. https://doi.org/10.1002/1099-095X(200011/12)11:6<745::AID-ENV431>3.0.CO;2-N

Cochran, W. G. (1952). The chi-square test of goodness of fit. *Annals of Mathematical Statistics*, *23*, 335−345.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*, 101−129.

Colling, L. J., & Szucs, D. (2018) Statistical inference and the replication crisis. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-018-0421-4

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*, 177–188. https://doi.org/10.1016/0197-2456(86)90046-2

de Souto Barreto, P., Rolland, Y., Vellas, B., & Maltais, M. (2019). Association of long-term exercise training with risk of falls, fractures, hospitalizations, and mortality in older adults: a systematic review and meta-analysis. *JAMA Internal Medicine*, *179*(3), 394–405. https://doi.org/10.1001/jamainternmed.2018.5406

de Vrieze, J. (2021). Large survey finds questionable research practices are common. *Science*, *373*(6552), 265. https://doi.org/10.1126/science.373.6552.265

Egger, M., Dickersin, K., & Davey Smith, G. (2001). Problems and limitations in conducting systematic reviews. In: Egger, M., Davey Smith, G., & Altman, D. G. (eds.) *Systematic reviews in health care: Meta−analysis in context, 2nd ed*. London, UK: BMJ Books.

Forstmeier, W., Wagenmakers, E. J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings – a practical guide. *Biological Reviews of the Cambridge Philosophical Society*, *92*(4), 1941−1968. https://doi.org/10.1111/brv.12315

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage Publications.

Glasziou, P., Meats, E., Heneghan, C., & Shepperd, S. (2008). What is missing from descriptions of treatment in trials and reviews? *British Medical Journal*, *336*(7659), 1472–1474. https://doi.org/10.1136/bmj.39590.732037.47

Greven, S., Dominici, F., & Zeger, S. (2011). An approach to the estimation of chronic air pollution effects using spatio-temporal information. *Journal of the American Statistical Association*, *106*(494), 396–406. https://doi.org/10.1198/jasa.2011.ap09392

Grimes, D. A., & Schulz, K. F. (2002). Cohort studies: marching towards outcomes. *Lancet 359*(9303), 341–345. https://doi.org/10.1016/S0140-6736(02)07500-1.

Harris, R. (2017). *Rigor mortis: How sloppy science creates worthless cures, crushes hope, and wastes billions*. New York, NY: Basic Books.

Hartshorne, J. K., & Schachner, A. (2012). Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Neuroscience*, *6*, 8. https://doi.org/10.3389/fncom.2012.00008

Hayat, M. J., Powell, A., Johnson, T., & Cadwell, B. L. (2017). Statistical methods used in the public health literature and implications for training of public health professionals. *PLoS ONE, 12(6)*, e0179032. https://doi.org/10.1371/journal.pone.0179032

Higgins, J. P. T., & Green, S. (eds). (2011). *Cochrane handbook for systematic reviews of interventions, version 5.1.0* [updated March 2011]. The Cochrane Collaboration, 2008. https://handbook-5-1.cochrane.org/

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, 327(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Hubbard, R. (2015). *Corrupt research: The case for reconceptualizing empirical management and social science*. London, UK: Sage Publications.

Hung, H. M. J., O'Neill, R. T., Bauer, P., & Kohne, K. (1997). The behavior of the p-value when the alternative hypothesis is true. *Biometrics, 53*, 11–22. https://doi.org/10.2307/2533093

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association, 294*, 218–228. https://doi.org/10.1001/jama.294.2.218

Ioannidis, J. P. A. (2008). Interpretation of tests of heterogeneity and bias in meta-analysis. *Journal of Evaluation in Clinical Practice, 14*(5), 951–957. https://doi.org/10.1111/j.1365-2753.2008.00986.x

Johnson, V. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, *110*(48), 19313−19317. https://doi.org/10.1073/pnas.1313476110

Kavvoura, F. K., Liberopoulos, G., & Ioannidis, J. P. (2007). Selection in reported epidemiological risks: An empirical

assessment. *PLoS Medicine*, *4*(3), e79. https://doi.org/10.1371/journal.pmed.0040079

Kindzierski, W., Young, S., Meyer, T., & Dunn, J. (2021). Evaluation of a meta-analysis of ambient air quality as a risk factor for asthma exacerbation. *Journal of Respiration*, *1*(3), 173−196. https://doi.org/10.3390/jor1030017

Lee, P. N., Forey, B. A., & Coombs, K. J. (2012). Systematic review with meta-analysis of the epidemiological evidence in the 1900s relating smoking to lung cancer. *BMC Cancer*, *12*, 385. https://doi.org/10.1186/1471-2407-12-385

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistical wars*. Cambridge, UK: Cambridge University Press.

McDonnell, W. F., Abbey, D. E., Nishino, N., & Lebowitz, M. D. (1999). Longterm ambient ozone concentration and the incidence of asthma in nonsmoking adults: the AHSMOG study. *Environmental Research*, *80*, 110–121. https://doi.org/10.1006/enrs.1998.3894

Morgenstern, V., Zutavern, A., Cyrys, J., Brockow, I., Gehring, U., Koletzko, S., Bauer, C. P., Reinhardt, D., Wichmann, H. E., & Heinrich, J. (2007). Respiratory health and individual estimated exposure to traffic related air pollutants in a cohort of young children. *Occupational and Environmental Medicine*, *64*, 8–16. https://doi.org/10.1136/oem.2006.028241

National Academies of Sciences, Engineering, and Medicine (NASEM). (1991). *Environmental epidemiology, Volume 1: Public health and hazardous wastes*. Washington, DC: The National Academies Press. https://doi.org/10.17226/1802

National Academies of Sciences, Engineering, and Medicine (NASEM). (2016). *Statistical challenges in assessing and fostering the reproducibility of scientific results: Summary of a workshop*. Washington, DC: The National Academies Press. https://doi.org/10.17226/21915

National Academies of Sciences, Engineering, and Medicine (NASEM). (2019). *Reproducibility and replicability in science*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303

Nissen, S. B., Magidson, T., Gross, K., & Bergstrom, C.T. (2016). Publication bias and the canonization of false facts. *eLife*, *5*, e21451. https://doi.org/10.7554/eLife.21451

Obenchain, R. L., & Young, S. S. (2017). Local Control strategy: Simple analyses of air pollution data can reveal heterogeneity in longevity outcomes. *Risk Analysis*, *37*, 1742−1753. https://doi.org/10.1111/risa.12749

Packer, M. (2017). Are meta-analyses a form of medical fake news? Thoughts about how they should contribute to medical science and practice. *Circulation*, *136*(22), 2097–2099. https://doi.org/10.1161/CIRCULATIONAHA.117.030209

Pearson, H. (2016). *The Life Project: The Extraordinary Story of Ordinary Lives*. London, UK: Allen Lane.

Pekkanen, J., & Pearce, N. (1999). Defining asthma in epidemiological studies. *European Respiratory Journal*, *14*, 951–957. https://doi.org/10.1034/j.1399-3003.1999.14d37.x

Pereira, T. V., & Ioannidis, J. P. (2011). Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. *Journal of Clinical Epidemiology*, *64*(10), 1060–1069. https://doi.org/10.1016/j.jclinepi.2010.12.012

Pocock, S. J., Collier, T. J., Dandreo, K. J., De Stavola, B. L., Goldman, M. B., Kalish, L. A., … McCormack, V. A. (2004). Issues in the reporting of epidemiological studies: A survey of recent practice. *British Medical Journal*, *329*, 883−888. https://doi.org/10.1136/bmj.38250.571088.55

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*(9), 712. https://doi.org/10.1038/nrd3439-c1

Randall, D., & Welser, C. (2018). *The irreproducibility crisis of modern science: Causes, consequences, and the road to reform*. New York, NY: National Association of Scholars. https://www.nas.org/reports/the-irreproducibility-crisis-of-modern-science/full-report

Ritchie, S. (2020). *Science fictions: How fraud, bias, negligence, and hype undermine the search for truth*. New York, NY: Henry Holt and Company.

Ryan, P. B., Madigan, D., Stang, P. E., Schuemie, M. J., & Hripcsak, G. (2013). Medication-wide association studies. *Pharmacometrics and Systems Pharmacology*, *2*, e76. https://doi.org/10.1038/psp.2013.52

Ravn, T., & Sørensen, M. P. (2021). Exploring the gray area: Similarities and differences in questionable research practices (QRPs) across main areas of research. *Science and Engineering Ethics*, *27*(4), 40. https://doi.org/10.1007/s11948-021-00310-z

Samet, J. M., & Muñoz, A. 1998. Evolution of the cohort study. *Epidemiologic Reviews, 20*(1), 1–14. https://doi.org/10.1093/oxfordjournals.epirev.a017964

Sarewitz, D. (2012). Beware the creeping cracks of bias. *Nature, 485*, 149. https://doi.org/10.1038/485149a

Schisterman, E. F., Cole, S. R., & Platt, R. W. (2009). Overadjustment bias and unnecessary adjustment in epidemiology studies. *Epidemiology*, *20*(4), 488–495. https://doi.org/10.1097/EDE.0b013e3181a819a1

Schnatter, A. R., Chen, M., DeVilbiss, E. A., Lewis, R. J., & Gallagher, E. M. (2018). Systematic review and meta-analysis of selected cancers in petroleum refinery workers. *Journal of Occupational and Environmental Medicine*, *60*(7), e329−e342. https://doi.org/10.1097/JOM.0000000000001336

Schweder, T., & Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, *69*, 493−502. https://doi.org/10.1093/biomet/69.3.493

Selwyn, M. R. (1989). Dual controls, *p*-value plots, and the multiple testing issue in carcinogenicity studies. *Environmental Health Perspectives, 82*, 337–344, https://doi.org/10.1289/ehp.8982337

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534−547. https://doi.org/10.1037/a0033242

Song, J. W., & Chung, K. C. (2010). Observational studies: cohort and case-control studies. *Plastic and Reconstructive Surgery*, *126*(6), 2234–2242. https://doi.org/10.1097/PRS.0b013e3181f44abc

Sterne, J. A. C., Egger, M., & Davey Smith, G. (2001). Investigating and dealing with publication and other biases in meta-analysis. *British Medical Journal*, *323*, 101–105. https://doi.org/10.1136/bmj.323.7304.101

Streiner, D. L. (2018). Statistics commentary series, commentary No. 27: p-hacking. *Journal of Clinical Psychopharmacology*, *38*(4), 286−288. https://doi.org/10.1097/JCP.0000000000000901

Vasilevsky, N. A., Brush, M. H., Paddock, H., Ponting, L., Tripathy, S. J., Larocca, G. M., & Haendel, M. A. (2013). On the reproducibility of science: Unique identification of research resources in the biomedical literature. *PeerJ, 1*, e148. https://doi.org/10.7717/peerj.148

Wang, X., & Yin, L. (2013). Identification of confounding versus dispersing covariates by confounding influence. *Communications in Statistics—Theory and Methods*, *42*, 4540–4556. https://doi.org/10.1080/03610926.2011.650267

Ware, M., & Mabe, M. (2015). *The STM Report: An overview of scientific and scholarly journal publishing*. Lincoln, NB: University of Nebraska. http://digitalcommons.unl.edu/scholcom/9

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing*. New York, NY: John Wiley & Sons.

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Young, S. S., Bang, H., & Oktay, K. (2009). Cereal-induced gender selection? Most likely a multiple testing false positive. *Proceedings of the Royal Society B: Biological Sciences*, *276*, 1211−1212. https://doi.org/10.1098/rspb.2008.1405

Young, S. S. (2019). SOP for evaluation of a meta-analysis. Center for Open Science: Charlottesville, VA. https://osf.io/eh68y/

Young, S. S., Acharjee, M. K., & Das, K. (2019). The reliability of an environmental epidemiology meta-analysis, a case study. *Regulatory Toxicology and Pharmacology*, *102*, 47−52. https://doi.org/10.1016/j.yrtph.2018.12.013

Young, S. S., & Karr, A. (2011). Deming, data and observational studies. *Significance*, *8*(3), 116−120. https://doi.org/10.1111/j.1740-9713.2011.00506.x

Young, S. S., & Kindzierski, W. B. (2019). Evaluation of a meta-analysis of air quality and heart attacks, a case study. *Critical Reviews in Toxicology*, *49*(1), 85−94. https://doi.org/10.1080/10408444.2019.1576587

Young, S. S., Kindzierski, W. B., & Meyer, T. (2022). Understanding p-hacking and HARKing. https://researchers.one/articles/22.01.00007

Young, S. S., Smith, R. L., & Lopiano, K. K. (2017). Air quality and acute deaths in California, 2000−2012. *Regulatory Toxicology and Pharmacology*, *88*, 173−184. https://doi.org/10.1016/j.yrtph.2017.06.003

Appendix A

Appendix A.doc

Base papers used for search space counting are presented in the file “Appendix A.doc”.

Appendix B

Appendix B.doc

Three search space analysis examples are presented in the file “Appendix B.doc”.

Appendix C

Appendix C.doc

Summary statistics of datasets and p-value plots for true null associations and true effects between two variables in observational studies are presented in the file “Appendix C.doc”.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).