Regulatory Toxicology and Pharmacology 86 (2017) 177–180

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

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

Air quality environmental epidemiology studies are unreliable

S. Stanley Young

CGStat, 3401 Caldwell Drive, Raleigh, NC 27607-3326, United States

ARTICLE INFO

Article history:
Received 17 December 2016
Received in revised form
6 March 2017
Accepted 7 March 2017
Available online 8 March 2017

Keywords:
Environmental epidemiology
Air quality
Mortality
Observational studies
Multiple testing
Multiple modeling
Analysis search space

ABSTRACT

Ever since the London Great Smog of 1952 is estimated to have killed over 4000 people, scientists have studied the relationship between air quality and acute mortality. There are many hundreds of papers examining the question. There is a serious statistical problem with most of these papers. If there are many questions under consideration, and there is no adjustment for multiple testing or multiple modeling, then unadjusted p-values are totally unreliable making claims unreliable. Our idea is to determine the statistical reliability of eight papers published in Environmental Health Perspectives that were used in meta-analysis papers appearing in Lancet and JAMA. We counted the number of outcomes, air quality predictors, time lags and covariates examined in each paper. We estimate the multiplicity of questions that could be asked and the number of models that could be constructed. The results were that the median numbers of comparisons possible for multiplicity, models and search space were 135, 128, and 9568 respectively. Given the large search spaces, finding a small number of nominally significant results is not unusual at all. The claims in these eight papers are not statistically supported so these papers are unreliable as are the meta-analysis papers that use them.

1. Background and introduction

Epidemiology exhibits a notoriously poor record with a serious lack of reproducibility of published findings going back at least as far as Feinstein (1988) with continuing complaints: Taubes and Mann (1995), Ioannidis (2005), Kaplan et al. (2010), and Young and Karr (2011), to name only a few. Even the popular press is taking notice of the problems; Taubes (2007) and Hughes (2007) are two examples. See also Wikipedia (2017) Replication crisis. Ominously, there may be actual misuse and/or even deliberate abuse of model fitting methods; see Clyde (2000), Glaeser (2006), Young and Karr (2011). In 2002, Norman Breslow noted that students with the same training and the same data set produced statistical models with vastly different claims, Breslow (2003). In 2010, two groups of researchers using the same data base of observational data found that a treatment both caused, Cardwell et al. (2010), and did not cause, Green et al. (2010), cancer of the esophagus. A Nature survey reported that 90% of scientists responding said there is crisis in science: a serious, 52%, or minor, 38%, crisis, Baker (2016).

The state of science is bad enough that a consumer of a science paper should start with the premise that any claim made is more likely than not to be wrong (it will fail to replicate).

The current US Environmental Protection Agency, EPA, paradigm is that PM2.5 is *causal* of acute human deaths. The then head of the EPA, Lisa Jackson, said “Particulate matter causes premature death. It doesn't make you sick. It's directly causal to dying sooner than you should.” She went on to say “If we could reduce particulate matter to levels that are healthy we would have an identical impact to finding a cure for cancer.” Cancer causes ~570,000 deaths per year.

This report unapologetically takes the position that the current paradigm, air quality is a killer, is not supported by statistical analysis that take multiple testing and multiple modeling into account and claims made in these papers may not replicate. Papers supporting the current paradigm are many. Google Scholar, “air pollution, mortality”, returns over 900,000 hits; Schwartz et al. (2017) is typical. These studies are almost always associational studies, and of course, association is not proof of causation. To examine our claim that the EPA paradigm is wrong, we start with two recent meta-analysis papers that look at air quality and mortality effects, Nawrot et al. (2011), Mustafic et al. (2012), hereafter Lancet and JAMA. Eight of the base papers used in these meta-analysis studies were published in Environmental Health Perspective, EHP; we examine those papers. Our thesis is that these papers are statistically flawed and that they may be part of a publication bias.

E-mail address: genetree@bellsouth.net.

http://dx.doi.org/10.1016/j.yrtph.2017.03.009
0273-2300/

A major contribution of this research is to show that a seriously flawed analysis strategy is used in these eight EHP papers rendering claims made in these papers unsupported.

2. Methods

In randomized clinical trials, RCTs, there is very careful attention given to the statistical analysis. A statistical protocol is developed and agreed to by the interested parties, often a drug company and the US FDA, before the study starts. One of the major concerns is the control of statistical false positive results. Statistical, experimental and managerial strategies are employed to control the false positive rate. Often replication of a finding is required. Contrast a RCT with the typical environmental observational study, EO. Environmental epidemiology essentially has few, if any, analysis requirements. In an EO study, the researcher can modify the analysis as the data is examined. Multiple outcomes can be examined, multiple variables (air components) can be used as predictors. The analysis can be adjusted by putting multiple covariates into and out of the model. It is thought that effects can be due to events on prior days so different lags can be examined. For example, PM2.5 yesterday or the day before can cause deaths today. Seldom, if ever, is there a written, statistical protocol prior to examination of the data. With these factors (outcomes, predictors, covariates, lags), there is no standard analysis strategy. The strategy can be try-this-and-try-that. Our method is simple counting and computing the size of the available analysis space.

3. Results

In Table 1, we give the numbers of outcomes, predictors, lags and covariates, for each of the eight papers. Functions of these counts can be used to estimate the number of questions, models and search space available analysis. The product of outcomes, predictors and lags gives the number of questions at issue. For example, three outcomes (AllCause Deaths, heart attacks, and stroke) can be paired with six predictors (CO, NO2, SO2, PM2.5, PM10, ozone) to give 18 possible questions. The number of models is given by $2^{\text{Covariates}}$, taking the position that each covariate can be in the model or not. The search space is the number of questions times the number of models.

The median sizes of questions, models and search space are 135, 128, and 9568 respectively. See Table 2. None of the eight papers mention correcting for multiple testing or multiple modeling. All papers appear to test at the level of 0.05. Given the multiple testing and multiple modeling, none of these papers provide strong evidence for their claims. Any claim made could easily be due to chance, a false positive. Note that each of these eight papers should be examined separately for strength of evidence. They must stand on their own before they can be considered for combining in a meta-analysis. As the base papers do not appear reliable, the meta-analysis papers, Lancet and JAMA, also appear unreliable.

Table 2
Number of questions, models, and total search space, medians and quartiles.

	Median	25%	75%
Multiple Questions	135	66	168
Multiple Models	128	20	448
Total Search Space	9568	2920	40,704

4. Discussion

There are many ways to increase the number of analysis options beyond our simple counting. We count two genders and two possible analyses (gender is in the analysis or not), but the analysis could be male, female and combined giving three options. In examination of a dose response, mortality versus PM2.5 level, logistic regression could be used, one model. Doing a transformation of the dose, say log, points the way to trying multiple transformations, Ginevan and Watkins (2010). Often the dose is cut into several groups, which offers further opportunities for model searching. Age can be treated as a continuous variable or cut into groups with an analysis in each group. Mann et al. (2003) do an analysis for each of three age groups so it could enter the counting process as three rather than two, in or out of the model. Temperature is obviously cyclical. It can be treated in any of several ways. Temperature effects can be controlled by use of a spline curve with differing degrees of stiffness. Or analysis can be within seasons. If case crossover analysis is used, comparisons are often within a month. Each of the analysis options could be changed from outcome to outcome and differ for each of the air components. Multi-component models could be computed, e.g. PM2.5 and ozone together in a model. These various methods could be explored giving the analyst many options for analysis.

After the dramatic increase in deaths after the Great London Smog, there was considerable search for the causative agent. The current paradigm, PM2.5 is a killer, essentially starts with Dockery et al. (1993). That paper now has over 7000 citations. In effect, their association claim is usually taken that PM2.5 is causative of deaths. The dramatic claim of Dockery fell upon very fertile ground. Dockery has been much criticized; the data set has been examined, but it is not publicly available.

Arguably a contemporaneous study was better, Styer et al. (1995). The sample size was much larger and the statistical analysis was sound. They tried a wide range of models and they found no consistent air quality effect on mortality. That paper is cited only just over 100 times. Both Dockery and Styer were funded by EPA.

The positive Dockery paper was take as valid and became the operational paradigm. Once a new paradigm is accepted (in this case by the EPA), it is expected that scientists will come in to fill in the gaps, Kuhn (1962), (and take advantage of funding opportunities). Subsequently many positive association studies were published. An editor commented to me, “The issue addresses (sic) was laid to rest in the mid 1990s by a large reanalysis report sponsored

Table 1
Counts of questions at issue in eight Environmental Health Effect papers used in two meta-analysis papers.

Reference	Outcomes	Predictors	Lags	Covariates	Questions	Models	Search Space
1 Koken et al., 2003	5	6	5	5	150	32	4800
2 Linn et al., 2000	10	4	3	7	120	128	15,360
3 Mann et al., 2003	4	4	6	9	96	512	49,152
4 Ye et al., 2001	16	7	5	3	560	8	4480
5 Zanobetti and Schwartz, 2005	1	1	3	7	3	128	384
6 Rich et al., 2010	5	5	7	10	175	1024	179,200
7 Zanobetti and Schwartz, 2009	5	6	5	4	150	16	2400
8 Barnett et al., 2006	7	4	2	8	56	256	14,336

by HEI. EPA and other regulatory bodies have long since concluded these associations are causal so I don't think there is much point in going over this again and again." in rejecting one of my papers without review.

It is rather routine for editors to reject negative studies out of hand. Informal conversations with multiple authors of published negative studies support the difficulty of getting them published. For example, there is evidence that Environmental Health Perspectives has a policy of rejecting negative papers. If they have that policy, they are not alone. Across the board, negative studies have a more difficult time getting published. Eventually we can have serious publication bias, positive studies are accepted as they support the current paradigm and negative studies are rejected. So far as we know observational studies used in meta-analyses are not routinely examined for multiple testing and multiple modeling bias. For more discussion of publication bias see Wikipedia (2017), Publication bias.

There is something of an art to writing of a scientific paper. Humans like a good story. The positive is accentuated and facts that do not fit are downplayed or even omitted, Glaeser (2006). Consider three marker negative papers, Styer et al. (1995), 115 citations, Chay et al. (2003), 103 citations and Enstrom (2005), 62 citations. Styer is cited only once in the eight papers and then not fairly. Chay is not cited in any of the four papers published after 2003. Enstrom is not cited in any of the three papers published after 2005. Schwartz et al. (2017) does not cite any of the three marker negative papers nor the important Greven et al. (2011) negative paper. In general, paradigm-negative papers are not cited by paradigm positive papers.

The primary author of each of the eight base papers was contacted twice asking if analysis data set used in their paper was available. None of the authors provided their analysis data set. Without access to the analysis data sets it is not possible to adjust the analysis for multiple testing and multiple modeling. From what is available in the base papers, it appears that none of the claims made in the eight papers would be statistically significant after adjustment.

It is not possible to prove a negative so to make a claim, an investigator should provide strong evidence, an analysis that names all the questions at issue and fairly adjusts for multiple testing and multiple modeling. None of the claims made in these EHP papers can be considered reliable due to inadequate analysis. The data should be made public so that the analysis can be corrected for multiple testing and multiple modeling.

A necessary requirement for numbers coming from a base paper to be combined in a meta-analysis is that the numbers be unbiased estimates of the quantity at issue, Boos and Stefanski (2013). The numbers can vary by chance from the target quantity, but they cannot be biased.

We, the science community, are letting the authors get away with doing exploratory data analysis repeatedly. They look at multiple outcomes, multiple causes, any number of covariates, and any number of time lags. They try this and try that and publish a paper if they get a p-value less than 0.05 where a plausible story can be made. If they fail to find "statistical significance," then it appears that they simply do not publish, creating publication bias. Authors, editors and consumers can become true believers in a false paradigm.

Here is a missing insight. In real science, a hypothesis is refined, and then retested with new data on a sharp question. The protocol is written before the new data is analyzed. There is statistical error control. There is replication. Logically the results of the new, more definitive study should take precedence over the exploratory studies. If it is positive, we say the hypotheses is supported. Popper, pure and simple. If the new study fails, we should say the hypothesis fails and spend science resources on some other problem.

It is very easy for humans to become true believers, especially when there is funding. Those doing air quality and health effects research should be held to good scientific standards. See Kabat (2017) pages 51–55.

5. Summary

Eight papers from Environmental Health Perspectives used in one or both meta-analysis studies were carefully examined with respect to the range of analysis options open to the researcher, the size of the analysis search space. The search space for each paper is large (in many cases vast) so that testing claims at a nominal 0.05 level is problematic. Any meta-analysis using these papers should also be considered unreliable until the reliability of the underlying papers is assured.

6. Next steps

It is recommended that the editor of Environmental Health Perspectives mark the eight papers as "Exploratory Study, not to be used for decision making". As the meta-analysis papers are not reliable, the editors of Lancet and JAMA should consider marking them "Withdrawn until the base papers are corrected for bias and a new meta-analysis is done."

Funding

This work was partially supported by the American Petroleum Institute.

Conflict of interest

I have no conflict of interests.

Transparency document

Transparency document related to this article can be found online at http://dx.doi.org/10.1016/j.yrtph.2017.03.009.

References

Eight environmental health perspectives references

Barnett, A.G., Williams, G.M., Schwartz, J., et al., 2006. The effects of air pollution on hospitalizations for cardiovascular disease in elderly people in Australian and New Zealand cities. Environ. Health Perspect. 114, 1018–1023.

Koken, P.J., Piver, W.T., Ye, F., Elixhauser, A., Olsen, L.M., Portier, C.J., 2003. Temperature, air pollution, and hospitalization for cardiovascular diseases among elderly people in Denver. Environ. Health Perspect. 111, 1312–1317.

Linn, W.S., Szlachcic, Y., Gong Jr., H., Kinney, P.L., Berhane, K.T., 2000. Air pollution and daily hospital admissions in metropolitan Los Angeles. Environ. Health Perspect. 108, 427–434.

Mann, J.K., Tager, I.B., Lurmann, F., et al., 2003. Air pollution and hospital admissions for ischemic heart disease in persons with congestive heart failure or arrhythmia. Environ. Health Perspect. 110, 1247–1252.

Rich, D.Q., Kipen, H.M., Zhang, J., Kamat, L., Wilson, A.C., Kostis, J.B., 2010. Triggering of transmural infarctions, but not nontransmural infarctions, by ambient fine particles. Environ. Health Perspect. 118, 1229–1234.

Ye, F., Piver, W.T., Ando, M., Portier, C.J., 2001. Effects of temperature and air pollutants on cardiovascular and respiratory diseases for males and females older than 65 years of age in Tokyo. July and August 1980–1995 Environ. Health Perspect. 109, 355–359.

Zanobetti, A., Schwartz, J., 2005. The effect of particulate air pollution on emergency admissions for myocardial infarction: a multicity case-crossover analysis. Environ. Health Perspect. 113, 978–982.

Zanobetti, A., Schwartz, J., 2009. The effect of fine and coarse particulate air pollution on mortality: a national analysis. Environ. Health Perspect. 117, 898–903.

General references

Baker, M., 2016. 1,500 scientists lift the lid on reproducibility. Nature 533, 452–454. http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970.

Boos, D.D., Stefanski, L.A., 2013. Essential Statistical Inference: Theory and Methods. Springer Science+Business Media, New York. Page 188.

Breslow, N.E., 2003. Are statistical contributions to medicine undervalued? Biometrics 59, 1–8.

Cardwell, C.R., Abnet, C.C., Cantwell, M.M., Murry, L.J., 2010. Exposure to oral bisphosphonates and risk of exophageal cancer. J. Am. Med. Assoc. 304, 657–663.

Chay, K., Dobkin, C., Greenstone, M., 2003. The clean air act of 1970 and adult mortality. J. Risk Uncertain. 27, 279–300.

Clyde, M., 2000. Model uncertainty and health effect studies for particulate matter. Environmetrics 11, 745–763.

Dockery, D.W., Pope III, C.A., Xu, X., Spengler, J.D., Ware, J.H., Fay, M.E., et al., 1993. An association between air pollution and mortality in six U.S. cities. N. Engl. J. Med. 329, 1753–1759.

Enstrom, J.E., 2005. Fine particulate air pollution and total mortality among elderly Californians. 1973–2002. Inhal. Toxicol. 17, 803–816.

Feinstein, A.R., 1988. Scientific standards in epidemiologic studies of the menace of daily life. Science 242, 1257–1263.

Ginevan, M.E., Watkins, D.K., 2010. Logarithmic dose transformation in epidemiologic dose-response analysis: use with caution. Regul. Toxicol. Pharmacol. 58, 336–340.

Glaeser, E.L., 2006. Researcher Incentives and Empirical Methods. Discussion paper 2122. http://scholar.harvard.edu/files/glaeser/files/researcher_incentives_and_empirical_methods.pdf [Last accessed on November 21, 2016].

Green, J., Czanner, G., Reeves, J., Watson, Wise, L., Beral, V., 2010. Oral bisphosphonates and risk of cancer of oesophagus, stomach, colerectum: case-control analysis with a UK primary care cohort. Br. Med. J. 341, c4444.

Greven, S., Dominici, F., Zeger, S., 2011. An approach to the estimation of chronic air pollution effects using spatio-temporal information. J. Amer Stat. Assoc. 106, 396–406.

Hughes, S., 2007. New York Times Magazine Focuses on Pitfalls of Epidemiological Trials. September 18. http://www.medscape.com/viewarticle/789597.

Ioannidis, J.P.A., 2005. Contradicted and initially stronger effects in highly cited clinical research. J. Am. Med. Assoc. 294, 218–229.

Kabat, G.C., 2017. Getting Risk Right: Understanding the Science of Elusive Health Risks. Columbia University Press, NY.

Kaplan, S.H., Billimek, J., Sorkin, D.H., Ngo-Metzger, Q., Greenfield, S., 2010. Who can respond to treatment? Identifying patient characteristics related to heterogeneity of treatment effects. Med. Care 48, S9–S16.

Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press.

Mustafic, H., Jabre, P., Caussin, C., Murad, M.H., Escolano, S., Tafflet, M., Perier, M.-C., Marijon, E., Vernerey, D., Empana, J.-P., Jouven, X., 2012. Main air pollutants and myocardial infarction: a systematic review and meta-analysis. J. Am. Med. Assoc. 307, 713–712.

Nawrot, T.S., Perez, L., Künzli, N., Munters, E., Nemery, B., 2011. Public health importance of triggers of myocardial infarction: a comparative risk assessment. Lancet 377, 732–740.

Schwartz, J., Bind, M.A., Koutrakis, P., 2017. Estimating causal effects of local air pollution on daily deaths: effect of low levels. Environ. Health Perspect. 125, 23–29.

Styer, P., McMillan, N., Gao, F., Davis, J., Sacks, J., 1995. Effect of outdoor airborne particulate matter on daily death counts. Environ. Health Perspect. 103, 490–497.

Taubes, G., 2007. Do We Really Know what Makes Us Healthy? Times Magazine, New York. September 16. http://www.nytimes.com/2007/09/16/magazine/16epidemiology-t.html.

Taubes, G., Mann, C.C., 1995. Epidemiology faces its limits. Science 269, 164–169.

Wikipedia. Publication bias. 2017. https://en.wikipedia.org/wiki/Publication_bias.

Wikipedia. Replication crisis. 2017. https://en.wikipedia.org/wiki/Replication_crisis.

Young, S.S., Karr, A., 2011. Deming, data and observational studies: a process out of control and needing fixing. Significance 8, 122–126 (September).