Multiple Testing and Multiple Modeling Problems with Observational Studies

S. Stanley Young [1], Warren Kindzierski [2]
[1]CGStat, 3401 Caldwell Drive, Raleigh, NC 27607
[2]University of Alberta, 3-57 South Academic Building,
11405-87 Avenue, Edmonton, Alberta, T6G 1C9 Canada

Abstract

There is a replication problem in science. Some of the causes are presented. We focus on multiple testing and multiple modeling, MTMM, in this paper. It is impossible to have severe testing if there is no control over MTMM. We focus on observational studies although there is good evidence of problems with experimental studies. A meta-analysis uses statistics coming from base papers to examine/make a claim. We start with a claim coming from a meta-analysis. We use two strategies to evaluate the claim. First, we examine the base papers and count the number of questions at issue. Second, we plot the ranked p-values from the base papers against the integers, 1, 2, …, N, a p-value plot. If the p-values are predominantly small, say <0.05, the claim is supported. If they form a 45-degree line, then the claim is not supported. We usually see a surprising result: a hockey stick figure. The small p-values on the blade of the hockey stick imply a real effect. The large p-values on the handle of the hockey stick imply no effect.

W. Edwards Deming would say that asking workers to fix a system where they are successful is doomed. He would identify the replication problem as a management problem, not a worker problem.

Key Words: Multiple testing, multiple modeling, reliability of literature claims, meta-analysis, p-value plot

1. Background and Introduction

Clearly society has derived much benefit from science and technology, but scientists themselves are saying there is a crisis of claims not replicating, Baker 2016, with 52% saying the crisis is major and 38% saying minor. Out of 52 observational study claims tested in randomized trials, none replicated in the expected direction and five were statistically significant but in the unexpected direction, Young and Karr, 2011. See Table 1. Even claims coming from experimental studies are not faring well. Of 53 claims coming from experimental biology, 47 could not be usefully replicated, Begley and Ellis 2012. Of 100 claims coming from experimental psychology only 36 replicated, Open Science Collaboration 2015. All the studies that failed to replicate came with statistical justification. What went wrong? Statistical methods, when carefully employed, work as expected, e.g. drug company randomized clinical trials, industrial experimentation, Box, Hunter and Hunter 1978.

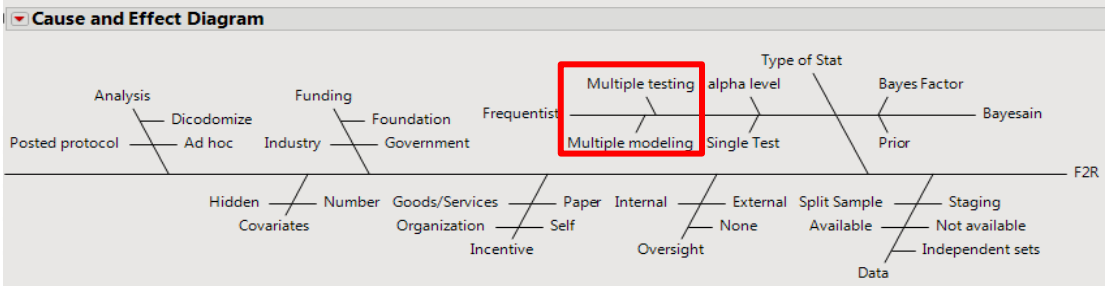
Table 1. Observational results tested in RCTs, Young and Karr 2011.

ID	Journal	Year	Treatment	Pos	Neg	#Studies

1	*NEJM*	1994	VitE, beta-carotene	0	1	3
2	*JAMA*	2003	HRT	0	3	4
3	*JNCI*	2005	VitE, beta-carotene	0	1	2
4	*JAMA*	2005	VitE	0	0	3
5	*JAMA*	2006	Low Fat	0	0	3
6	*NEJM*	2006	VitD, Calcium	0	0	3
7	*NEJM*	2006	Folic acid, Vit B6, B12	0	0	2
8	*JAMA*	2007	Low Fat	0	0	2
9	*AIntMed*	2007	VitC, VitE, beta-carotene	0	0	12
10	*JAMA*	2008	VitC, VitE	0	0	12
11	*JAMA*	2009	VitE, Selenium	0	0	3
12	*JAMA*	2002	HRT + Vitamins	0	0	3

To help understand process failures, researchers can use a cause-and-effect-diagram, Ishikawa 1968. There is a failure-to-replicate, F2R, of published literature. The possible causes are given in Figure 1.

Figure 1. A cause and effect diagram for failure to replicate, F2R. Given on the left are the possible causes for F2R.



There are many causes exhibited. The relative importance of the causes should be investigated. In this paper we choose to focus on multiple testing and multiple modeling, MTMM, taking the position if many tests are computed and many models are explored that false results can occur by chance. Even the blind hog occasionally finds an acorn.

2. **Methods**

We start our investigation with the examination of meta-analysis papers. In a meta-analysis, the researcher gathers papers that address a single question. This start has the important advantage of replication. The standard analysis used in a meta-analysis is to take statistics from the base papers, usually a risk ratio and its confidence limits, and combine these statistics to get a better estimate of the effect of interest. Replication allows comparisons among the papers.

Our first step is to simply count the number of questions as issue in each of the base papers. We count the number of outcomes, the number of predictors and the number of covariates. These can be combined to give an estimate of the analysis search space, Young 2017, Young and Kindzierski 2019. Examples of these counts are given in

Results. Obviously if the search space is large, the researcher has an opportunity to select a result for presentation. If there is no correction for MTMM, then there is an increased risk of false positive results.

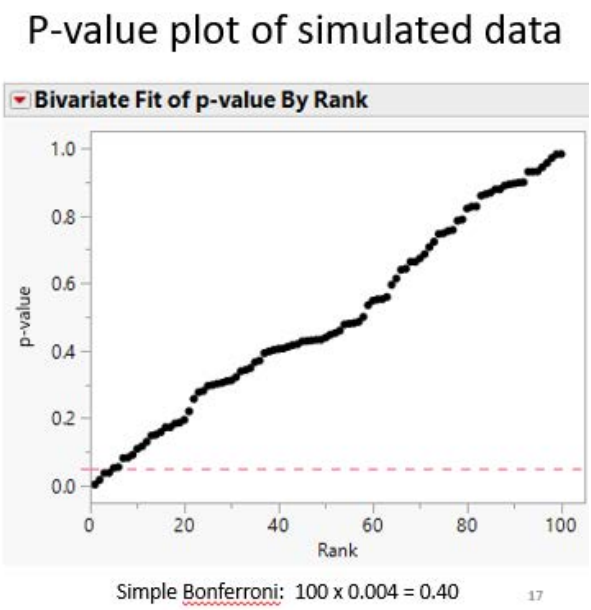
Our second step is to take advantage of the replication across base studies. For each risk ratio and its confidence limits, we compute a p-value. We then make a p-value plot as follows. The p-values are ranked from smallest to largest and plotted against the integers, 1, 2, 3, …, N. We slightly modify the results of Schweder and Spjøtvoll 1982, and follow the method used in Westfall and Young 1993. The expected interpretation is straightforward: if the p-values fall roughly on a 45-degree line, there is no effect; If most of the p-values are less than 0.05, then there is a treatment effect. We have noted the rather surprising result of a hockey stick pattern! Some of the p-values are small implying a real effect, while others fall on a 45-degree line, implying no effect. Examples will be given in Results.

3. **Results**

Our first result is a simple simulation. Three 10-sided dice can be cast to get red/white/blue digits to simulate a p-value. In a workshop, dice assure the participants that the process is random. Here we used a uniform random number generator to get numbers that ranged from 0.000 to 1.000. We give 100 p-values presented as 20 rows (tumors) and five columns (countries).
Figure 2. 100 simulated p-values and a p-value plot.

Figure 2. Simulated p-values and a p-value plot.

	Tumor	1	2	3	4	5
1	T01	0.899	0.417	0.673	0.754	0.686
2	T02	0.299	0.349	0.944	0.405	0.878
3	T03	0.868	0.535	0.448	0.430	0.221
4	T04	0.439	0.897	0.930	0.500	0.257
5	T05	0.429	0.082	0.038	0.478	0.053
6	T06	0.432	0.305	0.056	0.403	0.821
7	T07	0.982	0.707	0.460	0.789	0.956
8	T08	0.723	0.931	0.827	0.296	0.758
9	T09	0.174	0.982	0.277	0.970	0.366
10	T10	0.117	0.339	0.281	0.746	0.419
11	T11	0.433	0.640	0.313	0.310	0.482
12	T12	0.004	0.412	0.428	0.195	0.184
13	T13	0.663	0.552	0.893	0.084	0.827
14	T14	0.785	0.398	0.895	0.393	0.092
15	T15	0.595	0.322	0.159	0.407	0.663
16	T16	0.553	0.173	0.452	0.859	0.889
17	T17	0.748	0.480	0.486	0.018	0.130
18	T18	0.643	0.371	0.303	0.614	0.149
19	T19	0.878	0.548	0.038	0.864	0.152
20	T20	0.559	0.343	0.187	0.109	0.930



Simulated p-values less that 0.05 are colored. There are four marked p-values. One is rather small, 0.004. The p-values were ranked and plotted against the integers 1, 2, 3, ….100 to give a p-value plot. We observe a roughly 45-degree line.

3.1 Counting

We turn to counting results. Eight papers were published in the journal Environmental Health Perspectives (Impact factor 10.08) and used in a meta-analysis published in the

Journal of the American Medical Association (Impact factor 14.78). The counts and estimated search spaces are given in Table 2.

Table 2. The number of questions is outcomes x predictors. The number of models is 2 to the power of number of covariates. The search Space is the number of Questions times the number of Models.

RowID	Author	Year	Questions	Models	Space
1	Zanobetti	2005	3	128	384
2	Zanobetti	2009	150	16	2,400
3	Ye	2001	560	8	4,480
4	Koken	2003	150	32	4,800
5	Barnett	2006	56	256	14,336
6	Linn	2000	120	128	15,360
7	Mann	2003	96	512	49,152
8	Rich	2010	175	1024	179,200

The median search space (number of questions x number of models) is 9,568 with an interquartile range of 2,920 to 40,704.

There are many cohort studies reported in the literature. In a cohort study, people are examined initially and then followed over time. Somewhat surprisingly a single cohort study can give rise to thousands of papers. It can be and often is mix and match. There are multiple measurements at each time point and any predictor variable at any time point might show an association with a disease. Any striking association can be reported. We examined an environmental epidemiology meta-analysis that used cohort studies. The meta-analysis reported that the cohort studies supported a claim that poor air quality was associated with lung cancer. To get some idea of the wide-ranging nature of the claims coming from these studies, we searched literature for the number of papers that used or cited specific named cohort data base. Table 3 gives the number of papers in literature that each cohort data base we searched is used or cited.

Table 3 The number of papers found derived from data sets used in a meta-analysis of cohort studies

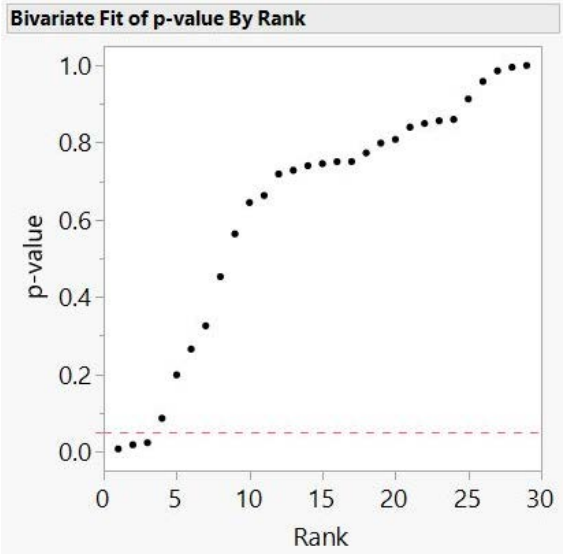
# papers/ citations	Cohort
1960	Netherlands Cohort Study on Diet and Cancer
98	China National Hypertension Survey
8380	Clinical Practice Research Datalink
195	Rome Longitudinal Study
32	Trucking Industry Particle Study
454	National Enhanced Cancer Surveillance System
111	Three-prefecture Cohort Study
7890	Cancer Prevention Study II
3530	Harvard Six Cities Study
3800	California Teachers Study
731	AHSMOG
1460	Nurses' Health Study (in title)
1940	European Study of Cohorts for Air Pollution effects

Clearly, these data bases are analyzed for multiple questions. None of these papers that we examined do any correction for MTMM. In no sense, should any claim coming from these data bases be considered anything but exploratory.

3.1 P-value plots

We now present some examples of p-value plots. First, we show a no effect study. Barreto et al. 2018 examine the effects of exercise programs on old people. They examined six outcomes: risk of falls, injurious falls, multiple falls, fractures, hospitalization, and mortality. Among their claims they found no effect of exercise programs on overall mortality, Figure 3

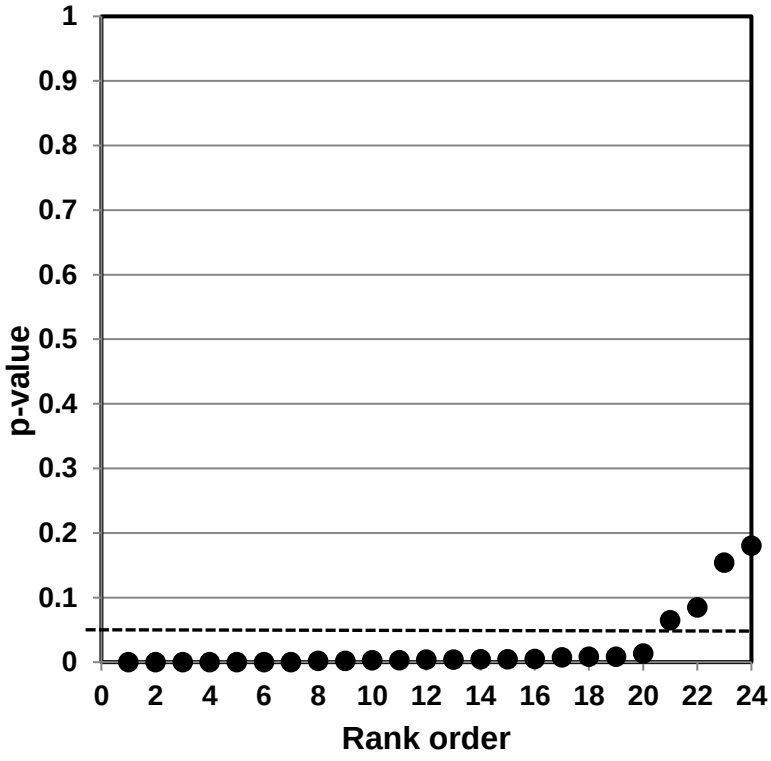
Figure 3. Mortality p-value plot for 29 studies, Barreto et al. 2018.



We see the mortality p-values for 29 studies. The standard random effects meta-analysis declares no effect. The p-value plot has an unusual appearance but indicates no effect also.

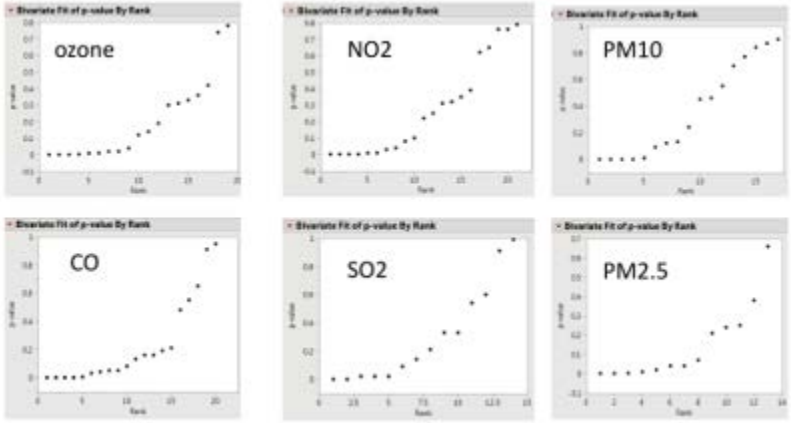
We next turn to a p-value plot of a positive effect. In a meta-analysis, Marsh et al. 2017 studied pleural malignant mesothelioma (PMM) risk among people exposed to asbestos non-occupationally. Twenty-four studies were found. We converted the RR and CLs to p-values and a p-value plot is given in Figure 4. We see that only a few of the p-values are greater than 0.05 and those are small as well.

Figure 4. P-value plot based on Marsh et al. 2017 base studies.



Our final p-value plots come from Young and Kindzierski 2019. A highly cited JAMA 2012 meta-analysis (542 citations as of Sept 13, 2020) examined six air components and claimed that all but ozone were associated with heart attacks, Mustafic 2012.

Figure 5. Six p-value plots presented in **Young and Kindzierski 2019**.



All six of the air component p-value plots have a hockey stick appearance with variable numbers of p-values on the blade or the handle.

We offer some thoughts on how the replication problem, including the MTMM problem, can be fixed. W. Edwards Deming would say that asking workers to fix a system where they are successful is doomed. He would identify the replication problem as a management problem, not a worker problem. System managers – funding agencies and journal editors – have responsibility for instituting reform. Among the reforms, funding agencies should support oversight: a. fund the building of data sets separately from their analysis; b. no funding without data and analysis code made public; and c. fund replication studies and re-analysis proposals. Journal editors should: a. label each study they publish as Exploratory or Confirmatory, and b. judge papers on protocol, data, and methods, not results. Finally, consumers should ignore all claims until management fixes the problem, i.e. assures severe testing.

4. Conclusions

Failure to replicate is a serious problem in science. All the features given in Figure 1 deserve some attention. Multiple testing and multiple modeling, MTMM, seem common and there are sound ways to deal with that problem, e.g., see Westfall and Young 1993. Funding agencies and editors should have accountability in laying down the law on researchers and helping fix the MTMM problem. Effectively dealing with multiple testing and multiple modeling would be a big step in the direction of sever testing, Mayo 2018.

Our suggestion of estimating the analysis search space and doing a p-value plot are easily accomplished, and they offer defensible insights into the reliability of a literature claim. Here we assert that MTMM is a problem for environmental epidemiology. Our experience is that it is difficult to find a p-value plot that does not have a hockey stick shape for environmental epidemiology meta-analysis data.

Subject experts should examine their own science areas.

Acknowledgements

Dr. Young received financial support from the National Association of Scholars.

References

Baker, M. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533, 452e454. http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Barreto PdS, Rolland Y, Vellas B, Maltais M. 2018. Association of long-term exercise training with risk of falls, fractures, hospitalizations, and mortality in older adults: A systematic review and meta-analysis. *JAMA Intern Med*. doi:10.1001/jamainternmed.2018.5406

Begley CG, Ellis LM. 2012. Drug development: Raise standards for preclinical cancer research. *Nature* 483, 531-533. http://www.nature.com/nature/journal/v483/n7391/full/483531a.html?foxtrotcallback=true.

Box GEP, Hunter WG, Hunter JS. 1978. Statistics for Experimenters. John Wiley & Sons. New York, New York

Ishikawa, K. 1968. *Guide to Quality Control. Tokyo: JUSE.*

Mayo, DG. 2018. *Statistical Inference as Severe Testing. Cambridge University Press. Cambridge, UK.*

Marsh GM, Riordan AS, Keeton KA, Benson SM. 2017. Non-occupational exposure to asbestos and risk of pleural mesothelioma: review and meta-analysis. *Occup Environ Med* 74, 838–846.

Mustafic H, Jabre P, Caussin C, Murad MH, 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. *JAMA*. 307,713-721. doi:10.1001/jama.2012.126.

Open Science Collaboration 2015. Estimating the reproducibility of psychological science. *Science 349, 4716.* DOI: 10.1126/science.aac4716

Schweder T, Spjøtvoll E. 1982. Plots of p-values to evaluate many tests simultaneously. *Biometrika*. 69:493-502.

Westfall PH, Young SS. 1993. *Resampling-based Multiple Testing*. John Wiley & Sons. New York, NY

Young SS. 2017. Air quality environmental epidemiology studies are unreliable. *Regulatory Toxicology and Pharmacology* 88, 177-180.

Young SS, Karr A. 2011. Deming, data, and observational studies. *Significance*. 8, 116-120. doi:10.1111/j.1740-9713.2011.00506.x.

Young SS, Kindzierski WB. 2018. Background information for meta-analysis evaluation, Info 01-Info 06. https://arxiv.org/abs/1808.04408.

Young SS, Kindzierski KB. 2019. Evaluation of a meta-analysis of air quality and heart attacks, a case study. *Critical Reviews in Toxicology*. doi:10.1080/10408444.2019.1576587