Commentary

Karl E. Peace¹ / JingJing Yin¹ / Haresh Rochani¹ / Sarbesh Pandeya¹ / Stanley Young²

A Serious Flaw in Nutrition Epidemiology: A Meta-Analysis Study

¹ Jiann Ping-Hsu College of Public Health, Georgia Southern University, Statesboro, GA 30458, USA, E-mail:

kepeace@georgiasouthern.edu

² CGSTAT, Raleigh, NC 27607, USA**Abstract:**

Background: Many researchers have studied the relationship between diet and health. Specifically, there are papers showing an association between the consumption of sugar sweetened beverages and Type 2 diabetes. Many meta-analyses use individual studies that do not attempt to adjust for multiple testing or multiple modeling. Hence the claims reported in a meta-analysis paper may be unreliable as the base papers do not ensure unbiased statistics.

Objective: Determine (i) the statistical reliability of 10 papers and (ii) indirectly the reliability of the meta-analysis study.

Method: We obtained copies of each of the 10 papers used in a metaanalysis paper and counted the numbers of outcomes, predictors, and covariates. We estimate the size of the potential analysis search space available to the authors of these papers; i. e. the number of comparisons and models available. The potential analysis search space is the number of outcomes times the number of predictors times 2^c , where c is the number of covariates. This formula was applied to information found in the abstracts (Space A) as well as the text (Space T) of each base paper.

Results: The median and range of the number of comparisons possible across the base papers are 6.5 and (2 12,288), respectively for Space A, and 196,608 and (3072–117,117,952), respectively for Space T. It is noted that the median of 6.5 for Space A may be misleading as each study has 60–165 foods that could be predictors.

Conclusion: Given that testing is at the 5% level and the number of comparisons is very large, nominal statistical significance is very weak support for a claim. The claims in these papers are not statistically supported and hence are unreliable so the meta-analysis paper is also unreliable.

Keywords: observational studies, nutritional epidemiology, reliability of claims, multiple testing, multiple modeling, meta-analysis

DOI: 10.1515/ijb-2018-0079

Received: August 10, 2018; Revised: October 31, 2018; Accepted: November 4, 2018

1 Background

In a meta-analysis, a statistic, e. g. a mean, a ratio, etc., is taken from each base paper and combined, e. g. averaged, weighted averaged, to give a better measure of some finding. (See Chen and Peace [1] and Ehm [2]) Key requirements of statistics coming from base papers used in a meta-analysis are that the statistics are independent and unbiased, Boos and Stefanski [3].

For a meta-analysis of randomized studies, these conditions are usually met; the papers are independent and randomization gives unbiased estimates. These conditions may not be met for a meta-analysis of observational studies and that is the subject of this paper.

Epidemiology exhibits a notoriously poor record of reproducibility of published findings going back at least as far as Feinstein [4], and Mayes et al. [5], in 1988, with continuing complaints: Taubes and Mann [6], Ioannidis [7], Kaplan et al. [8], Young and Deming [9], and Breslow [10, 11], to name a few. Breslow commented that “contradictory results emanating from a plethora of irreproducible observational studies have contributed to the lack of esteem with which epidemiology is regarded by many in the wider biomedical community.” Even the popular press is speaking up; Taubes [12], and Hughes [13], are two examples. See also Wikipedia [14], Replication crisis. Ominously, there may be actual misuse and/or even deliberate abuse of model fitting methods; see Glaeser

Karl E. Peace is the corresponding author.

© 2018 Walter de Gruyter GmbH, Berlin/Boston.

[15], Young and Deming [9]. Two groups of researchers using the same observational data base found that a treatment both caused, Cardwell et al. [16], and did not cause, Green et al. [17], cancer of the esophagus.

A Nature survey reported that 90% of scientists responding said there is a serious (52%) or minor (38%) crisis in science, Baker [18]. The state of published scientific claims is sufficiently suspect that a consumer of information from such publications should start with the presumption that any claim made is as likely as not to be wrong (it will fail to replicate).

In randomized clinical trials (RCTs), very careful attention is given to the statistical analysis. A protocol, with a statistical analysis section (SAS) and a statistical analysis plan (SAP), is developed and agreed to by the interested parties, often a drug company and the Food Drug Administration (FDA), before the study starts. One of the major concerns is the control of false positive results, a biased answer. Statistical, experimental and managerial strategies (identified in the SAP) are employed to minimize the occurrence of a false positive result. Often replication of a finding is required.

Contrast a RCT with the typical nutritional observational study. Nutritional epidemiology essentially has no analysis requirements. Typically, the researcher can modify the analysis strategy as the data is examined. Multiple outcomes can be examined and multiple variables can be used as predictors. The analysis can be adjusted by putting multiple covariates into and out of the model. Seldom, if ever, is there a written protocol. For these factors (outcomes, predictors, covariates), there is no standard analysis strategy. The improvised strategy is often try-this-and-try-that. Under these circumstances the analysis is essentially exploratory.

This report asserts that many claims made on the basis of meta-analyses of nutrition or diet and certain diseases have not been proven as the underlying papers are exploratory. Many papers exist that examine the role of sugar sweetened beverages, SSBs, in the development of related chronic metabolic diseases, such as metabolic syndrome and type 2 diabetes, the question examined by Malik et al. [19]. Google Scholar (12/22/2016) returns over 1,510,000 hits for association between diet or nutrition and diabetes; 1,290,000 hits for association between diet or nutrition and Type II diabetes; and 14,500 hits for association between consumption of sugar-sweetened beverages and Type II diabetes. These studies are almost always associational studies, and of course, association is not proof of causation. Malik et al. [19], with a meta-analysis looks at sugar-sweetened beverages and the risk of metabolic syndrome and type 2 diabetes, hereafter referred to as Malik. We examine the 10 papers [20–29] upon which the metaanalysis is based. Our thesis is that these papers are essentially exploratory and as confirmatory papers they are statistically flawed. In addition, there may be publication bias.

A major contribution of this research is to show that the analysis strategy used in the 10 base papers produce biased statistics which are unsuitable for meta-analysis.

2 Methods

A protocol and data extraction form (DEF) were developed (Appendix) and the methods therein followed.

2.1 Screening and evaluation methods

We read the meta-analysis and base papers, filled in the data extraction form (DEF), and requested raw data from the lead author of each base paper. Data extracted were: sample sizes, pvalues, relative risks, confidence limits outcomes, predictors, covariates and funding sources. Functions of these counts were used to estimate the potential size of the analysis search space available to the researcher; i. e. the number of comparisons and models available.

The potential analysis search space for each base paper was computed as follow

$$\text{Search Space} = \text{No. of Outcome} \times \text{No. of Predictors} \times 2^c$$

where c is the number of covariates. This formula was applied to information found in both the abstract (Space A) as well as the text (Space T) of each base paper.

2.2 Operation

Two teams were formed, each consisting of an Assistant Professor of Biostatistics, a DrPH student and a Master's level student. Membership of the teams was determined randomly.

The 10 base papers were randomly assigned in balanced fashion to the two teams. Each team carefully reviewed the assigned 5 papers and extracted data therefrom in period 1. The 5 papers reviewed by Team 1 were then crossed over to Team 2 for review and data extraction and vice versa. Differences in extraction results between the two teams were resolved by the Co-PIs. A final DEF was completed for each paper. All final DEFs were posted to the study folder on the Google drive in PDF format. All final DEFs are available as supplemental material.

3 Results

Across the 10 studies (Table 1), sample sizes ranged from 4,304 to 91,249, with a median of 28,897 and a total of 332,357; smallest nominal p-values ranged from 0.0001–0.001; and largest reported relative risk (RR) ranged from 1.23–5.06 with a median of 2.07. Eighty percent (80%) of the studies reported only government funding, 10% reported both government and nongovernment funding and 10% was unfunded.

Table 1: Review of sample size for 10 base papers.

Paper ID	Overall Sample Size	Sample Size per Group
Nettleton et al. [20]	6,814	Rare or never: 2961, > rare/never but < 1 servings per week: 455, ≥1 servings/week to < 1 servings/day: 914, ≥1 serving/day: 681
Lutsey et al. [21]	9,514	Men: 4197, Women: 5317
Dhingra et al. [22]	8,997	<1 soft drinks per day: 5840, 1 soft drinks per day: 1918, ≥2 soft drinks per day: 1239
Montonen et al. [23]	4,304	1st quartile: 1076, 2nd quartile: 1076, 3rd quartile: 1076, 4th quartile: 1076
Paynter et al. [24]	12,204	Men: 5414, Women: 6790
Schulze et al. [25]	91,249	For 1991, <1/mo: 49,203, 1–4/mo: 23,398, 2–6/wk: 9950, <1/d: 8698; For 1991–1995, ≤1/wk: 38,737, ≥1/d: 2366, ≤1/wk to ≥1/d: 1007, ≥1/d tp ≤ 1/wk: 1020
Palmer et al. [26]	43,960	Soft drinks per week: <1: 25,971, 2–6: 10,521, ≥1: 7468; Fruit Drinks per Week : <1: 15,455, 2–6: 13,722, ≥ 1: 13,644
Bazzano et al. [27]	71,346	quintile 1: 14,573, quintile 2: 14,408, quintile 3: 14,337, quintile 4: 14,118, quintile 5: 13,913
Odegaard et al. [28]	43,580	Soft drink consumption: almost never: 32,060, 1–3/Month: 4514, 1/week: 2389, 2–3/week: 4617; Juice Consumption: almost never: 35,719, 1–3/Month: 4399, 1/week: 1791, 2–3/week: 1671
de Koning et al. [29]	40,389	Sugar Sweetened beverages: Q1: 13,675, Q2: 5022, Q3: 11,729, Q4: 9963; Artificially sweetened beverages: Q1: 18,442, Q2: 2681, Q3: 9448, Q4: 9818
Across all articles	332,357	

The number of outcomes, predictors and covariates for each of the 10 base papers appear in Table 2. The range and median of the number of comparisons possible across the base papers are (2 12,288) and 6.5, respectively for Space A (Table 3), and (3072 117,117,952) and 196,608, respectively for Space T (Table 4). None of the 10 papers mention correcting for multiple testing or multiple modeling or adjusting for multiplicities of any kind. All papers appear to test at the 5% level.

Table 2: P-values, relative risks, multiplicity adjustment & funding source for 10 base papers.

Paper ID	Smallest p-value	Largest RR (Hazard ratio)	Largest RR: CI	Multiplicity Adjustment for p-values	Funding Source
Nettleton et al. [20]	<0.001	2.2	(1.1–4.51)	No	Government
Lutsey et al. [21]	<0.001	1.34	(1.24–1.44)	No	Government
Dhingra et al. [22]	<0.0001	2.31	(1.77–3.01)	No	Government and Non-government
Montonen et al. [23]	<0.001	5.06	(1.87–3.71)	No	Unfunded
Paynter et al. [24]	<0.01	1.23	(0.93–1.62)	No	Government

Schulze et al. [25]	<0.001	2.31	(1.55–3.45)	No	Government
Palmer et al. [26]	0.001	1.51	(1.31–1.75)	No	Government
Bazzano et al. [27]	<0.001	4.47	(2.35–7.66)	No	Government
Odegaard et al. [28]	<0.0001	1.7	(1.34–2.16)	No	Government
de Koning et al. [29]	<0.01	1.94	(1.75–2.14)	No	Government
Across all articles	<0.0001→<0.01	1.23–5.06			90% Government

Table 3: Search space size of 10 base papers based on Abstracts.

Base Papers	Base Papers Journals	Outcomes	Predictors	Covariates	Space Size
Nettleton et al.	Diabetes Care [20]	2	1	3	16
Lutsey et al.	Circulation [21]	1	2	4	32
Dhingra et al.	Circulation [22]	7	1	10	7,168
Montonen et al.	J Nutr [23]	1	5	0	5
Paynter et al.	Am J Epidem [24]	1	2	0	2
Schulze et al.	JAMA [25]	2	1	2	8
Palmer et al.	Arch Intern Med [26]	1	1	2	4
Bazzano et al.	Diabetes Care [27]	1	3	0	3
Odegaard et al.	Am J Epidem [28]	1	2	2	8
de Koning et al.	Am J Epidem [29]	1	3	10	12,288

Table 4: Space size of 10 base papers based on Texts of Papers.

Base Papers	Base Papers Journals	Outcomes	Predictors	Covariates	Space Size
Nettleton et al.	Diabetes Care [20]	2	2	15	196,608
Lutsey et al.	Circulation [21]	1	2	14	32,678
Dhingra et al.	Circulation [22]	7	1	24	117,117,952
Montonen et al.	J Nutr [23]	1	12	15	392,396
Paynter et al.	Am J Epidem [24]	1	2	14	32,678
Schulze et al.	JAMA [25]	2	3	9 [Mod 1]	3,072
Palmer et al.	Arch Intern Med [26]	2	3	15	196,608
Bazzano et al.	Diabetes Care [27]	1	5	13	40,960
Odegaard et al.	Am J Epidem [28]	2	2	16	262,144
de Koning et al.	Am J Epidem [29]	1	3	24	6,291,456

As it is impossible to “prove a negative,” it is the responsibility of a researcher making a claim to provide strong evidence in support of the presumed positive claim. Given the multiple testing and multiple modeling, none of these papers provide strong evidence for their claims. Any claim made could easily be a false positive. Note that each of these 10 papers should be examined separately for validity of inferences. They must stand on their own before they can be considered for combining in a meta-analysis. As the statistics used in the base papers do not provide valid evidence for claim, the validity of the claim from the meta-analysis paper is questionable.

It is useful to review order statistics, their expected values and their relation to expected p-values as a function of the number of observations in a sample. If a random sample is taken from a population, and the objects are ordered from smallest to largest, the reordered objects are called order statistics. The value of the largest order statistic in the sample does not change from its value in the unordered sample, but it is a different animal. It is the largest number in the sample. The larger the sample, the larger is the expected value of the largest object (see Table 5). Consider a sample from the normal distribution with a standard deviation of one. If there are 10 objects in the sample, then the expected value of the largest object is 1.54.

Table 5: The expected value* of largest order statistics $E(X_{(n)})$ and corresponding P-values for a sample of set size N from a standard normal distribution (i. e. $Z \sim N(0,1)$).

N	Exp. Value of largest Order Statistic	P-Value
10	1.53875	0.12211
20	1.86748	0.06976

30	2.04276	0.04952
40	2.16078	0.03864
50	2.24907	0.03181
60	2.31928	0.02709
70	2.37736	0.02364
80	2.42677	0.02099
90	2.46970	0.01890
100	2.50759	0.01720
125	2.58634	0.01407
150	2.64925	0.01194
175	2.70148	0.01038
200	2.74604	0.00919
225	2.78485	0.00826
250	2.81918	0.00750
300	2.87777	0.00635
350	2.92651	0.00551
400	2.96818	0.00487
1000	3.24144	0.00119
5000	3.67755	0.00024

* The expected value is calculated by using the pdf of order statistics in equation (5.4.4) by the book Statistical Inference from Casella and Berger [36]. The p-values is calculated by $P(|Z| \geq E(X_{(n)}))$, which can be interpreted as under the null, by pure chance, the unadjusted p-value that we often use to reject null by comparing it with a nominal significance level. Therefore, under large sample size, the small values of unadjusted p-values can be meaningless.

Besides the expected value, the p-value for a z-test against the value zero is given. We expect the largest value in a sample of 10 to be about 0.5 standard deviations from the mean. Now look down Table 4. As N increases, the expected value of the largest order statistic increases. In a sample of 30 we expect the largest order statistic, by chance alone, to be about 2.04 standard deviations larger than the mean (of zero). The corresponding (unadjusted) p-value is 0.0495, which would be nominally statistically significant.

It is statistically fatal to consider the largest order statistic as if it were a random observation! It is common to adjust p-values when there are many questions at issue to control the false positive error rate. This table can be used to remind a researcher that the value of an object can be large by chance alone and that a p-value can be small, again by chance alone. The value of the largest order statistics will be larger, the larger the sample and the p-value will be smaller, the larger the sample.

If, after adjustment, a p-value is not statistically significant, the researcher needs to keep in mind that the corresponding experimental value is an order statistic and needs to be judged by expected values of order statistics, not as if it were a random value from the distribution in question.

The researcher can “cut” a continuous variable to create ordered groups. The low group can be used as a reference group and the other groups can be compared to the reference group. The set of groups can be tested for linear trend. In Table 6, the number of p-values displayed in each paper is given as # tests.

Table 6: Risk ratios, confidence limits taken from Table 1 of [19]. Z-tests, p-values, adjustment factors and adjusted p-values were computed.

Ref	Sig	RR	CLL	CLH	Beta	BetaSE	Z	Prob	AdjFactor	AdjP
Nettleton et al.	0.05	0.86	0.62	1.17	−0.151	0.162	−0.931	0.8241	116-736	1.000
Lutsey et al.	p-val	1.09	0.99	1.19	0.086	0.047	1.836	0.0332	540-672	1.000
Dhingra et al.	CL 95%	1.39	1.21	1.59	0.329	0.070	4.726	<0.0001	244	0.000
Montonen et al.	CL 95%	1.67	0.98	2.87	0.513	0.274	1.871	0.0307	102-400	1.000
Paynter et al.	CL 95%	1.17	0.92	1.39	0.122	0.157	1.491	0.0679	244	1.000
Schulze et al.	0.05	1.83	1.42	2.36	0.604	0.130	4.663	<0.0001	217-9072	1.000
Palmer et al.	CL 95%	1.24	1.06	1.45	0.215	0.080	2.692	0.0036	222-8224	1.000
Bazzano et al.	CL 95%	1.31	0.99	1.74	0.270	0.144	1.877	0.0303	112-64	1.000
Odegaard et al.	CL 95%	1.42	1.25	1.62	0.351	0.066	5.301	<0.0001	135-1680	0.078
de Koning et al.	0.05	1.14	1.03	1.28	0.131	0.055	2.364	0.0090	8384	1.000
Nettleton et al.*	0.05	1.15	0.92	1.42	0.140	0.111	1.262	0.1034	116-736	1.000

*This was not included in the pool of ten base papers that but was reported by Malik et al. [19].

Table 7: Number of foods, FFQ, considered in each base paper, the total number of covariates, the number of groupings used for predictors and the type of statistical testing – based on Table 1 of Malik. Also given are the number of p-values reported, #Tests, derived by counting in each base paper.

Ref	FFQ	Total	#groups	Method	#Tests
Nettleton et al.	114	10	4	Trend, Each vs control	88
Lutsey et al.	66	13	5	Trend, Each vs control	85
Dhingra et al.	61	2	3	Trend, Each vs control	101
Montonen et al.	100	10	4	Trend, Each vs control	63
Paynter et al.	61	2	5	Trend, Each vs control	60
Schulze et al.	133	14	4	Trend, Each vs control	54
Palmer et al.	68	15	3	Trend, Each vs control	87
Bazzano et al.	88	7	5	Trend, Each vs control	114
Odegaard et al.	165	13	4	Trend, Each vs control	50
de Koning et al.	131	6	4	Trend, Each vs control	84
Nettleton et al.*	114	10	4	Trend, Each vs control	88

*This was not included in the pool of ten base papers that but was reported by Malik et al. [19].

Example R code is as follows:

```
# N=10 to 5000

n.dat=5000

f <-function(x, mu=0, sigma=1) dnorm(x, mean=mu, sd=sigma)
F <-function(x, mu=0, sigma=1) pnorm(x, mean=mu, sd=sigma,
lower.tail=FALSE)

#the pdf of X(r) of size n is given in Casella and Berger p 229 equation
(5.4.4)
integrand <-function(x,r,n,mu=0, sigma=1){
  x * (1 F(x, mu, sigma))^(r-1) * F(x, mu, sigma)^(n-r) * f(x, mu, sigma)
}
# the expectation is given as E(x)=integrate (x*pdf) from -inf to inf.
E <-function(r,n, mu=0, sigma=1){
  (1/beta(r,n-r+1)) * integrate(integrand,-Inf,Inf, r, n, mu, sigma)
  $value
}
E(n.dat,n.dat)

#the p-value is the probability being more extreme than the largest order
stat on the two-sided tails

2*(1-pnorm( E(n.dat,n.dat) ) )
```

To make things concrete, suppose that there are 60 questions at issue in a study, and to simplify this discussion, assume the questions are independent of one another, then by chance alone we would expect to see a mean value of 2.31 standard deviations and a p-value of 0.027 for the largest order statistics. Without taking order statistics and multiple testing into account, we would declare statistical significance AND we would not expect the result to replicate. We would have a false positive. Taking the value of 2.31 to a meta-analysis would be totally misleading. The classic reference is Royston [30], who also quotes a well-known formula by Blom [31].

Statistics from Table 1 of Malik et al. [19], were extracted and placed in our Table 5 and Table 6. Using their Risk Ratios and Confidence Limits we computed z-tests, unadjusted and adjusted p-values. Note that in our Table 5 and Table 6, we have two rows for Nettleton, one for diabetes and one for metabolic syndrome. After Bonferroni adjustment, there are no statistically significant results, which implies that the observed risk ratios are biased.

In our Table 6, we give some more characteristics related to statistical testing. We note the number of foods in the food questionnaires used, FFQ. The number varies from a low of 61 to a high of 165. Each of these foods could be used individually or in combination as a predictor of the health effect. The number of covariates given

explicitly in Table 2 of Malik et al. [19], is given as Total. Note that their counts of covariates are smaller than the number of covariates we counted (Table 4). Clearly the number of covariates mentioned in the abstract (Table 4), is an underestimate of the number of covariates in play.

It is common to group the predictors. In this case the number of groupings varies from 3 to 5 (Table 7). Using these groupings the researchers tested for a linear trend and they also tested the highest group against the lowest group (Table 7) so there were two dose response tests. We do not know if the number of groupings might have affected the results. Finally, each paper presents a large number of reported p-values, #Tests. None of the base papers reported any adjustment for multiple testing or multiple modeling.

4 Discussions

The first point to make is that the authors of the base paper were, in effect, doing exploratory analyses. The analysis search space for each paper was vast and nominal statistical significance of 5% is, at best, a screen, not confirmatory in any sense. A major multiple testing dilemma occurred in the 1990s when genomics came on line. Lander and Kruglyak [32], argued that for claims to be believable there should be multiple testing correction over all the analysis search space. None of the ten papers we examined performed any adjustment for multiple testing or multiple modeling and that appears to be usual for analysis of Food Frequency Questionnaires, FFQs.

Here is a missing insight. In real science, a hypothesis is refined, and then retested with new data on a sharper question. The protocol is written before the new data is analyzed. There is statistical error control. There is replication. We should give greater credence to the results of the new, more definitive study. If it is positive, we say the hypothesis is supported. If the new study fails, we should consider abandoning the hypothesis and spend science resources on some other problem.

If the covariates are fixed, then nutrition studies that use FFQs offer an opportunity for finding many negative findings. In FFQ studies there are approximately 60 to 130 food questions and many of these food questions are repeated from one study to another. The statistical analysis of all foods is easily accomplished with a few lines of code. A p-value plot would facilitate examination of all the questions, Schweder and Tvoll [33].

It is rather routine for a researcher not to submit negative papers as the belief is that editors are likely to reject negative papers. Informal conversations with multiple authors of published negative studies support the difficulty of getting them published. Across the board, negative studies have a more difficult time getting published. Given that negative papers are typically not published, eventually we can have serious publication bias, positive studies are accepted as they support the current paradigm and negative studies are rejected. As 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 [14], Publication bias.

Humans like a good story, which becomes a useful art in the writing of a scientific paper. Authors can accentuate positive papers and downplay or even omit negative papers, see Kabat [34]. It is very easy for presumptively neutral researchers to become true believers in an existing popular paradigm, especially when there is funding. Those doing nutrition and health effects research should be held to strict scientific standards: state if a study is exploratory, refine claims coming from an exploratory study for a confirmatory study, make data sets and analysis code available, etc.

Scientifically and logically, it is not possible to prove a negative so to make a public health 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 10 papers can be considered reliable due to potential bias, and hence they are inappropriate for inclusion in a meta-analysis.

We, the science community, are not recognizing that authors are doing exploratory data analysis over and over, year after year. They look at multiple outcomes, multiple causes, any number of covariates, and any number of predictors. They try this and try that analysis and publish a paper if they get a p-value less than 0.05 where a plausible story can be made Kass et al. [35]. If they fail to find "statistical significance," then it appears that they simply do not publish. Those doing meta-analyses need to realize the problem to their work. Authors, editors and consumers can become true believers in a false paradigm.

Finally, the primary author of each of the 10 papers was contacted twice asking if data used in their paper were available. None of the authors provided their analysis data set. Unfortunately, it is common for authors not to provide their analysis data set. Without access to the data sets it is not possible to adjust the analysis for multiple testing and multiple modeling. From what is available in the papers and as summarized in Table 1 of Malik [19], it appears that none of the claims made in the 10 papers would be statistically significant after adjustment. The data should be made public so that the analyses can be corrected for the bias introduced by multiple testing and multiple modeling.

5 Summary

Ten papers used in the meta-analysis study by Malik et al. [19] 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) in light of all the questions possible so that testing claims at a nominal 0.05 is problematic. Meta-analysis using these papers should also be considered unreliable until the reliability of the underlying papers is assessed or confirmatory studies are run.

Appendix

A Protocol Voz: MMA Study

Note: Sections of Protocol V01 were rewritten upon discovering that the Malik et al paper had only 10 non-overlapping base papers.

Co-PI: Karl Peace, Jiann-Ping Hsu College of Public Health, Georgia Southern University, kepeace@georgiasouthern.edu

Co-PI: Stan Young, CGStat, genetree@bellsouth.net

Background: For many nutritional questions randomized trials are not available so observational studies are conducted. It is common to gather a number of observational studies related to a question. The individual studies are evaluated and summary results from the studies are combined using what is called meta-analysis methods.

Idea: Our study is to evaluate the reliability of a nutritional meta-analysis study by examining the statistical reliability of the underlying studies.

The meta-analysis study of Malik et al. [19] was selected for study. Within the paper there appeared to be 11 cited base studies. However, upon examination of Dr Young, one appeared a replicate. Hence only the 10 nonoverlapping base papers were reviewed and contained data for extraction.

Objectives:

- Determine the size of the analysis search space for each observational base study of a meta-analysis.
- Determine if uncorrected summary statistics invalidate meta-analysis claims.

Study Population: Base papers from a meta-analysis paper of observational studies.

Locating studies: Reference list from the meta-analysis paper

Screening and Evaluation Methods:

- Read meta-analysis and base papers.
- Fill in Data Extraction Form.
- Ask for data access.

Operation:

Two teams will be formed, each consisting of an Assistant Professor of Biostatistics, a DrPH student and a Master's level student. Membership of the teams will be determined randomly.

The 10 base papers will be randomly assigned in balanced fashion to the two teams. Each team will review and extract data from the assigned 5 papers during period 1. The 5 papers reviewed by Team 1 will then be crossed over to Team 2 for review and data extraction and vice versa. Differences in extraction results between the two teams will be resolved by the Co-PIs. A final Data Extraction Form will be completed for each paper. All Data Extraction Forms will be posted to the study folder on the Google drive in PDF format.

The search space will be computed for each base paper as:

$\#outcomes \times \#predictors \times 2^c$, where c is the number of covariates in the final model.

Results: The summary results for a paper will be considered unreliable if the search space is greater than 100 or if $\#outcomes \times \#predictors$ is greater than 10. The meta-analysis paper will be considered unreliable if over $\frac{1}{4}$ if the base papers are considered unreliable.

References:

To minimize print space, the Malik et al paper is **reference** [19] and the 10 base papers are **references** [20–29] of the References section of the manuscript.

B Data Extraction Form Preliminary Final

MMA Study: Malik et al. Diabetes Care [19], 33, 2477–2483.

Your name: Date:

1. Paper (fill in the literature references as it appears in the meta-analysis paper)
2. PI: name, email address, regular mail
3. Journal editor: name, email address
- 4.A Overall Sample size:
- 4.B Sample size per Group (identify group)
 - Group 1: Sample Size
 - Group 2: Sample Size
 - Group 3: Sample Size
 - Group 4: Sample Size
5. Smallest p-value..... Largest RR with CL
6. # outcomes From Abstract From Paper
7. # predictors From Abstract From Paper
8. # covariates From Abstract From Paper
- 8.A # potential covariates mentioned
- 8.B # Covariates used in the analysis model
9. Is a food questionnaire used in the study? Yes No
10. Raw Data available (as stated in the paper)? Yes No
11. Funding source. Government Grant Number Industry Unfunded
12. Eligibility Criteria
13. Comments. Any other things of potential interest noted while reviewing the paper.

References

- [1] Chen DGD, Peace KE. *Applied meta-analysis with R*. Boca Raton: CRC Press, 2013.
- [2] Ehm W. Meta-analysis of mind-matter experiments: a statistical modeling perspective. *Mind Matter*. 2005;3:85–132.
- [3] Boos D, Stefanski L. Bayesian inference. *Essential statistical inference* Vol. 2013. In: EDs Boos D, Stefanski L (eds.). New York: Springer, 2013:163–203.
- [4] Feinstein A. Scientific standards in epidemiologic studies of the menace of daily life. *Science*. 1988;242:1247–64.
- [5] Mayes L, Horwitz R, Fhnstein A. A collection of 56 topics with contradictory results in case-control research. *Int J Epidemiol*. 1988;3:680–85.
- [6] Taubes G, Mann C. Epidemiology faces its limits. *Science*. 1995;269:164.
- [7] Ioannidis J. Contradicted and initially stronger effects in highly cited clinical research. *Jama*. 2005;2:218–28.
- [8] Kaplan S, Billimek J, Sorkin DH, Ngo-Metzger Q, Greenfield S. Who can respond to treatment?: identifying patient characteristics related to heterogeneity of treatment effects. *Med Care*. 2010;48:S9–S16.
- [9] Young S, Deming KA. Data and observational studies: a process out of control and needing fixing. *Significance*. 2011;8:116–20.
- [10] Breslow N. Are statistical contributions to medicine undervalued?. *Biometrics*. 2003;59:1–8.
- [11] Breslow N. Commentary. *Biostatistics*. 2010;3:379–80.
- [12] Taubes G. 2007. Do we really know what makes us healthy. New York Times, 2007.
- [13] Hughes S. 2007. New York times magazine focuses on pitfalls of epidemiological trials. 2007. <http://www.theheart.org/article/813719>.
- [14] Wikipedia. 2016. Replication Crisis. <https://en.wikipedia.org/wiki/Replicationcrisis>.
- [15] Glaeser E. 2006. Researcher incentives and empirical methods. National Bureau of Economic Research.
- [16] Cardwell C, Abnet C, Cantwell M, Murray LJ. Exposure to oral bisphosphonates and risk of esophageal cancer. *Jama*. 2010;304:657–63.

- [17] Green J, Czanner G, Reeves G, Watson J, Wise L, Beral V. Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: case-control analysis within a UK primary care cohort. *BMJ*. 2010;341:c4444.
- [18] Baker M. 1,500 scientists lift the lid on reproducibility. *Nature*. 2016;533:452–54.
- [19] Malik VS, Popkin B, Bray G, Despre' S J-P, Willett W, Hu F. Sugar-sweetened beverages and risk of metabolic syndrome and type 2 diabetes. *Diabetes Care*. 2010;33:2477–83.
- [20] Nettleton J, Lutsey P, Wang Y, Lima J, Michos E, Jacobs D. Diet soda intake and risk of incident metabolic syndrome and type 2 diabetes in the multi-ethnic study of atherosclerosis (MESA). *Diabetes Care*. 2009;32:688–94.
- [21] Lutsey P, Stevens J. Dietary intake and the development of the metabolic syndrome. *Diabetes Care*. 2008;117:754–61.
- [22] Dhingra R, Sullivan L, Jacques P, Wang T, Fox CS, Meigs J, et al. Soft drink consumption and risk of developing cardiometabolic risk factors and the metabolic syndrome. *Circulation*. 2007;116:480–88.
- [23] Montonen J, Ja'rvinen R, Knekt P, Helio'Vaara M, Reunanen A. Consumption of sweetened beverages and intakes of fructose and glucose predict type 2 diabetes occurrence. *J Nutr*. 2007;137:1447–54.
- [24] Paynter N, Yeh H-C, Voutilainen S, Schmidt M, Heiss G, Folsom A, et al. Coffee and sweetened beverage consumption and the risk of type 2 diabetes mellitus the atherosclerosis risk in communities study. *Am J Epidemiol*. 2006;164:1075–84.
- [25] Schulze M, Manson J, Ludwig DS, Colditz GA, Stampfer M, Willett W, et al. Sugar-sweetened beverages, weight gain, and incidence of type 2 diabetes in young and middle-aged women. *Jama*. 2004;292:927–34.
- [26] Palmer JR, Boggs D, Krishnan S, Hu F, Singer M, Rosenberg L. Sugar-sweetened beverages and incidence of type 2 diabetes mellitus in African American women. *Arch Intern Med*. 2008;164:1075–84.
- [27] Bazzano LA, Li T, Joshipura K, Hu F. Intake of fruit, vegetables, and fruit juices and risk of diabetes in women. *Diabetes Care*. 2008;31:1311–17.
- [28] Odegaard A, Koh W-P, Arakawa K, Mimi C, Pereira M. Drink and juice consumption and risk of physician-diagnosed incident type 2 diabetes the Singapore Chinese health study. *Am J Epidemiol*. 2010;171:701–8.
- [29] de Koning L, Malik V, Rimm E, Willett W, Hu FB. Sugarsweetened and artificially sweetened beverage consumption and risk of type 2 diabetes in men. *Am J Clin Nutr*. 2011;93:1321–27.
- [30] Royston J. Algorithm AS 177: expected normal order statistics (exact and approximate). *J R Stat Soc Ser C (Appl Stat)*. 1982;31:161–65.
- [31] Blom G. *Statistical estimates and transformed beta-variables*. Stockholm: Almqvist & Wiksell, 1958:174.
- [32] Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet*. 1995;11:241–47.
- [33] Schweder T, Tvoll ES. Plots of p-values to evaluate many tests simultaneously. *Biometrika*. 1982;3:493–502.
- [34] Kabat G. *Getting risk right: understanding the science of elusive health risks*. New York: Columbia University Press, 2016.
- [35] Kass R, Caffo B, Davidian M, Meng XL, Yu B, Reid N. Ten simple rules for effective statistical practice. *PLoS Comput Biol*. 2016;12(6):e1004961. DOI: 10.1371/journal.pcbi.1004961.
- [36] Berger R, Casella G. *Statistical Inference*Wadsworth statistics/probability series. Pacific Grove: Brooks/Cole Publishing Company, 1990.