International Journal of Statistics and Probability; Vol. 11, No. 3; May 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Statistical Reliability of a Diet-Disease Association Meta-analysis

S. Stanley Young[1] and Warren B. Kindzierski[2]

[1] CGStat, Raleigh, NC, USA

[2] Independent consultant, St Albert, Alberta, Canada

Correspondence: Warren B. Kindzierski, 12 Hart Place, St Albert, Alberta, T8N 5R1, Canada.

Received: April 4, 2022 Accepted: April 26, 2022 Online Published: April 27, 2022

doi:10.5539/ijsp.v11n3p40 URL: https://doi.org/10.5539/ijsp.v11n3p40

Abstract

Risk ratios or p-values from multiple, independent studies – observational or randomized – can be pooled to address a common research question in meta-analysis. However, reliability of independent studies should not be assumed as claimed risk factor−disease relationships may fail to reproduce. An independent evaluation was undertaken of a published meta-analysis of cohort studies examining diet−disease associations; specifically between red and processed meat and six disease outcomes (all-cause mortality, cardiovascular mortality, all cancer mortality, breast cancer incidence, colorectal cancer incidence, type 2 diabetes incidence). The number of hypotheses examined were counted in 15 random base papers (14%) of 105 used in the meta-analysis. Test statistics (relative risk values with 95% confidence limits) for 125 results used in the meta-analysis were converted to p-values; p-value plots were used to examine the effect heterogeneity of the p-values. The possible number of hypotheses examined in the 15 base papers was large, median = 20,736 (interquartile range = 1,728–331,776). Each p-value plot for selected health effects showed either a random pattern (p-values > 0.05), or a two-component mixture (small p-values < 0.001 while other p-values appeared random). Given potentially large numbers of hypotheses examined in the base studies, questionable research practices cannot be ruled out as explanations for some test statistics with small p-values. Like the original findings of the published meta-analysis, our independent evaluation concludes that base papers used in the meta-analysis do not support evidence for an association between red and processed meat and the six health effects investigated.

Keywords: cohort studies, red meat, health effects, meta-analysis, multiple testing bias

1. Introduction

Food and dietary intake habits represent a complex system of interacting components that may affect health status and disease over an individual's lifetime. Nutritional epidemiology uses methods to study how diet might affect health status and disease. These methodologies require a strong statistical component to develop useful and interpretable diet−disease associations (Prentice & Huang, 2018). The semi-quantitative food frequency questionnaire (FFQ) – a self-administered dietary assessment instrument – is commonly used to assess dietary intake (Boeing, 2013). A FFQ distributes a structured food list and a frequency response section to study participants, who indicate their usual frequency of intake of each food over a set period of time (Satija et al., 2015).

Causal criteria in nutritional epidemiology include (Potischman & Weed, 1999): consistency, strength of association, dose response, plausibility, and temporality. A longstanding critique of nutritional epidemiology in establishing causality is that it relies predominantly on observational study data, which researchers generally judge to be less reliable than experimental data (Satija et al., 2015). Bias – systematic alteration of research findings due to factors related to study design, data acquisition, statistical analysis, or reporting of results (Boffetta et al., 2008; NASEM, 2016, 2019; Randall & Welser, 2018) – can undermine a study's reliability to apply these causal criteria. Further, selective reporting occurs in published observational studies with researchers routinely testing many hypotheses during a study and then only reporting results that are interesting (i.e., statistically significant) (Gotzsche, 2006; Frieden, 2017).

One aspect of reproducibility – the performance of another study statistically confirming the same hypothesis or claim – is a cornerstone of science and reproducibility of research findings is needed before causal inference can be made (Moonesinghe et al., 2007). However, irreproducible published studies reportedly occur in a wide range of scientific disciplines – including general medicine, clinical sciences, oncology, nutrition, biology, psychological sciences (Young et al., 2022). Incomplete reporting occurs in biomedical research (Dickersin & Chalmers, 2011; Frieden, 2017). These types of situations can lead to an inability to reproduce research claims (Sarewitz, 2012). Part of the problem may arise from researchers examining large numbers of hypotheses and using multiple statistical models without statistical

correction – referred to as multiple testing and multiple modelling or multiple testing bias (Westfall & Young, 1993; Young & Kindzierski, 2019; Young et al., 2022).

Meta-analysis is a systematic procedure for statistically combining data (test statistics) from multiple studies that address a common research question (Egger et al., 2001), such as whether a particular food has an association with a disease. Meta-analysis has been placed at the top of the medical evidence-based pyramid – above case–control and cohort studies, and randomized trials (Murad et al., 2016). However, questions remain about whether the test statistics themselves being combined in meta-analysis may be derived using imperfect or limited statistical methodologies.

As a case in point, Peace et al. (2018) recently examined aspects of multiple testing associated with test statistics combined from ten base papers in a Malik et al. (2010) meta-analysis of sugar-sweetened beverage intake and risk of metabolic syndrome and type 2 diabetes. Peace et al. (2018) observed that none of the base papers in the Malik et al. meta-analysis corrected for multiple testing bias. Given the importance of statistics in developing useful and interpretable risk factor−disease associations, we were interested in understanding whether multiple testing bias might be occurring elsewhere in diet−disease association meta-analysis studies. Specifically, we randomly selected and independently evaluated base studies in a meta-analysis of the association between red and processed meat and selected human chronic effects.

2. Method

2.1 Data Sets

Vernooij et al. (2019) – herein referred to as Vernooij – published a meta-analysis of cohort studies relating to health claims from red and processed meat in the journal *Annals of Internal Medicine*. We selected six of 30 health effects that they examined for further independent evaluation – those that combined the largest number of base papers. These health effects included: all-cause mortality, cardiovascular mortality, all cancer mortality, breast cancer incidence, colorectal cancer incidence, type 2 diabetes incidence. Upon request, one of the Vernooij researchers provided data we used for our evaluation. We then used search space analysis (counting of the numbers of hypotheses examined in base studies) (Peace et al., 2018) and p-value plots (Schweder & Spjøtvoll, 1982) to evaluate the six diet−disease association claims.

Vernooij systematically reviewed 1,501 papers and selected 105 primary papers for further analysis. Their data set included 70 different population cohorts. They used GRADE (*Grading of Recommendations Assessment, Development and Evaluation*) criteria (Guyatt et al., 2008) – which do not assess multiple testing bias – to select base papers for their meta-analysis. Their study complied with recommendations of PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) (Moher et al., 2009).

Vernooij stated that the base papers used for meta-analysis, which were observational studies, provided low- or very-low-certainty evidence according to GRADE criteria. They concluded "…*dietary patterns with less red and processed meat intake may result in very small reductions in adverse cardiometabolic and cancer outcomes*." Numerous nutritional epidemiologists reacted to their research with some asking the editor of *Annals of Internal Medicine* to withdraw the paper before publication (Monaco, 2019; Arends, 2020).

2.2 Numbers of Hypotheses Tested in Single Studies (Counting)

One needs to estimate the number of hypotheses examined in a single study to assess the potential for multiple testing bias. We selected a subset of studies from Vernooij and counted the possible hypotheses examined in these studies. A 5 to 20% sample from a population whose characteristics are known is considered acceptable for most research purposes as it provides an ability to generalize for the population (Creswell, 2003). We believed the Vernooij judgment that their systematic review (screening) process selected 105 base papers with sufficiently consistent (known) characteristics for meta-analysis. We then randomly selected 15 of the 105 base papers (14%) for counting purposes.

The number of hypotheses considered in an individual base paper used by Vernooij was estimated as follows. Cohort studies generally use a direct statistical analysis strategy on data collected – e.g., what causes or risk factors are related to what outcomes (health effects). If a data set contains "C" causes and "O" outcomes, C × O possible hypotheses can be investigated. An adjustment factor "A" (also called a covariate) can be included as a yes/no adjustment – such as income or education – to see how it can modify each of the C × O hypotheses. Here an adjustment factor is included or excluded; and a multiplier of 2 is assumed for each adjustment factor considered. We counted causes (C), outcomes (O), and yes/no adjustment factors (A); where the number of hypotheses can be approximated as $= C \times O \times 2^{A}$.

We then specifically examined the 15 random base papers for evidence of whether a paper: i) mentioned multiple testing bias in different forms (i.e., multiple hypotheses or hypothesis, multiple testing, multiple comparisons, multiplicity) and/or, ii) made any mention of correcting for this bias.

2.3 P-value Plots

Epidemiologists traditionally use risk statistics (e.g., risk ratios or odds ratios) and confidence intervals instead of p-values from a hypothesis test to establish statistical significance. Given that researchers can estimate risk statistics, confidence intervals and p-values from the same data (Altman & Bland, 2011a,b), one can be estimated from the other. We estimated p-values from risk statistics and confidence intervals for all data used by Vernooij using JMP statistical software (SAS Institute, Cary, NC). We then developed p-value plots (Schweder & Spjøtvoll, 1982) to inspect the distribution of the set of p-values – i.e., the test statistics used by Vernooij.

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. In a well-designed and conducted study, the p-value is distributed uniformly over the interval 0 to 1 regardless of sample size under the null hypothesis and a distribution of true null hypothesis points plotted against their ranks in a p-value plot should form a 45-degree line when there are no effects (Schweder & Spjøtvoll, 1982; Hung et al., 1997; Bordewijk et al., 2020). Researchers can use the plot to assess the heterogeneity of the test statistics combined in meta-analyses.

The p-value plots were constructed and interpreted as follows:

- Computed p-values were ordered from smallest to largest and plotted against the integers, 1, 2, 3,…
- If p-value points on the plot followed an approximate 45-degree line, we concluded that test statistics resulted from a random (chance) process and the data supported the null hypothesis of no significant association.
- If p-value points on the plot followed approximately a line with a flat/shallow slope, where most (the majority) of p-values were small (< 0.05), then test statistic data set provided evidence for a real, statistically significant, association.
- If numbers of possible hypotheses tested were high in the base studies and p-value points on the plot exhibited a bilinear shape (divided into two lines), the data set of test statistics used for meta-analysis is consistent with a two-component mixture and a general (over-all) claim is not supported. In addition, a small p-value reported for the overall claim in the meta-analysis may not be valid (Schweder & Spjøtvoll, 1982).

Questionable research practices (QRP) involve approaches used by researchers during data collection, analysis, and reporting that may increase false-positive findings in published literature (Ware & Munafò, 2015; Kunert, 2016). P-value plotting is a useful tool to detect the possibility that QRP may have affected test statistics drawn into meta-analysis and rendered the meta-analysis unreliable.

2.4 Numbers of Hypotheses Tested on Cohort Population Data Sets

An interesting problem of multiple testing bias may exist with cohort population data sets. While it is time-consuming and expensive to set up and follow a new cohort, it can be relatively inexpensive to add new measurements and research questions (hypotheses) to an existing cohort. For these reasons, it is possible to have many hypotheses examined on a given cohort as data for the cohort can be used repeatedly. A single published study of a particular cohort data set may only address the tip of the iceberg in terms of numbers of hypotheses examined and multiple testing bias. Collectively there may be numerous other hypotheses at issue when one considers that the same cohort data set can be used many times over for research. Many published papers in literature based on a single cohort data set imply large number of hypotheses examined overall with the possibility of large numbers of false positive (chance) results reported in literature.

First, we wanted to show how common FFQ data is used by researchers investigating health effects. A potential problem is that researchers using FFQs – which are typically utilized in cohort studies – can examine many hypothesis and produce large numbers of false positive (chance) results. We did a Google Scholar (GS) database search to record the approximate number of articles in Web literature with the exact phrase "food frequency questionnaire" and a "[health effect]" mentioned anywhere in an article. We looked at 18 health effects: obesity, inflammation, depression, mental health, all-cause mortality, high blood pressure, lung and other cancers, metabolic disorders, low birth weight, pneumonia, autism, suicide, COPD (chronic obstructive pulmonary disease), ADHD (attention-deficit/hyperactivity disorder), miscarriage, atopic dermatitis, reproductive outcomes, erectile dysfunction.

Second, we did another GS database search to record the approximate number of articles in Web literature using food frequency questionnaire (FFQ) data for each cohort indicated in the 15 selected base papers from Section 2.2. We used the exact phrase "[cohort name]" and the term "FFQ" mentioned anywhere in an article for the search.

3. Results

3.1 Research Questions Asked in Single Studies (Counting)

Table 1 shows the count characteristics of 15 random papers selected from Vernooij. While early food frequency

questionnaire (FFQ) studies used only 61 foods (Willett et al., 1985), these 15 base papers include FFQ−cohort populations examining as many as 280 foods and 32 different health outcomes (Table 1). Summary statistics of the 15 base papers are presented in Table 2. The median number of causes (predictors) was 15 and the median number of adjustment factors (covariates) was 9 in Table 2. These numbers suggest a great scope of the numbers of hypotheses examined (search space).

Table 1. Characteristics of 15 randomly selected papers from Vernooij

Citation#	Base Paper 1st Author	Year	Foods	Outcomes	Causes (Predictors)	Adjustment Factors (Covariates)	Tests	Models	Search Space
8	Dixon	2004	51	3	51	17	153	131,072	20,054,016
31	McNaughton	2009	127	1	22	3	22	8	176
34	Panagiotakos	2009	156	3	15	11	45	2,048	92,160
38	Héroux	2010	18	32	18	9	576	512	294,912
47	Akbaraly	2013	127	5	4	5	20	32	640
48	Chan	2013	280	1	34	10	34	1,024	34,816
49	Chen	2013	39	4	12	5	48	32	1,536
53	Maruyama	2013	40	6	30	11	180	2,048	368,640
56	George	2014	122	3	20	13	60	8,192	491,520
57	Kumagai	2014	40	3	12	8	36	256	9,216
59	Pastorino	2016	45	1	10	6	10	64	640
65	Lacoppidan	2015	192	1	6	16	6	65,536	393,216
80	Lv	2017	12	3	27	8	81	256	20,736
92	Chang-Claude	2005	14	5	3	7	15	128	1,920
99	Tonstad	2013	130	1	4	10	4	1,024	4,096

Note: Citation# is Vernooij reference number; Author name is first author listed for reference; Year = publication year; Foods = # of foods used in Food Frequency Questionnaire; Tests = Outcomes × Causes; Models = 2^A where A = number of Adjustment Factors; Search Space = Tests × Models = approximation of number of hypotheses examined.

Researchers may believe they gain advantage by studying large numbers of outcomes, causes, and adjustment factors (i.e., testing many hypotheses), on the presumption that this maximizes their chances of discovering risk factor−health outcome associations (Willett et al., 1985). However, what they may have maximized is their likelihood of registering a false positive. Given that the conventional threshold for statistical significance in most disciplines is a p-value of less than 0.05, a false positive result should occur 5% of the time by chance alone in a multiple testing setting (Young et al., 2021). The median count of the 15 base papers was 20,736 (refer to Table 2). Five percent of 20,736 possible hypotheses examined in a single FFQ−cohort data set equals 1,037 chance findings that may be mistaken for real results.

Table 2. Characteristics of 15 randomly selected papers from Vernooij

Statistic	Foods	Outcomes	Causes (Predictors)	Adjustment Factors (Covariates)	Tests	Models	Search Space
minimum	12	1	3	3	4	8	176
lower quartile	40	1	8	7	18	96	1,728
median	51	3	15	9	36	512	20,736
upper quartile	129	5	25	11	71	2,048	331,776
maximum	280	32	51	17	576	131,072	20,054,016
mean	93	5	18	9	86	14,149	1,451,216

Note: Foods = # of foods used in Food Frequency Questionnaire; Tests = Outcomes × Causes; Models = 2^A where A = number of Adjustment Factors; Search Space = Tests × Models = approximation of number of hypotheses examined.

In our review of the 15 base papers for evidence of correction for multiple testing bias, thirteen of the papers made no mention of this bias. One paper (Panagiotakos et al., 2009) stated… '*multiple comparisons are made and consequently the probability of false positives findings (i.e., p-value) increases*'. Another paper (George et al., 2014) stated… '*All statistical tests were based on a priori hypotheses; therefore, no adjustment was performed for multiple testing*'.

However, the estimated search space (number of hypotheses examined) is > 490,000 for this paper (refer to Table 1). The only apparent a priori hypotheses stated in their paper were '*how scores on 4 commonly used diet quality indices – the Healthy Eating Index 2010, the Alternative Healthy Eating Index 2010, the Alternate Mediterranean Diet, and the Dietary Approaches to Stop Hypertension – are related to the risks of death from all causes, cardiovascular disease (CVD), and cancer among postmenopausal women*'.

3.2 P-value Plots

The p-value plots for six health outcomes are presented in Figure 1. Each of the six images in Figure 1 indexes rank order (the x axis) and p-value (the y axis). The p values – symbols (circles or triangles) in the body of the six images – are ordered from smallest to largest. The number of p-values in each plot corresponds to the number of studies (base papers) for each of the six outcomes. As noted in the Methods, if there is no effect the p-values will form roughly a 45º line. If the line is flat/shallow with most of the p-values small, then it supports a real effect. Finally, if the shape of the points is bilinear and the counts are high, then the result, i.e., claim, is ambiguous (uncertain) at best.

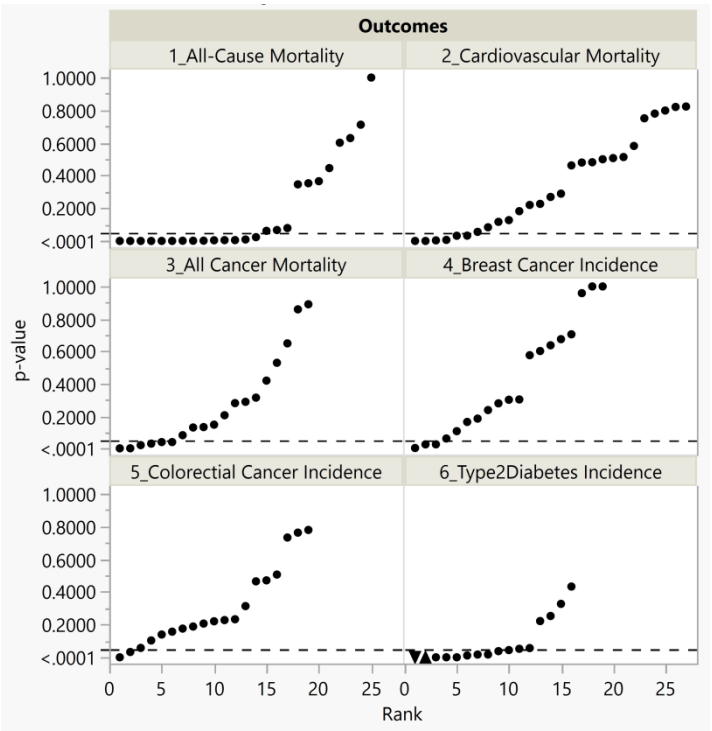

Figure 1. P-value plots (p-value versus rank) for meta-analysis of six health outcomes from Vernooij. Symbols (circles or triangles) are p-values ordered from smallest to largest; triangle pointing downwards (upwards) represents decreasing (increasing) effect

The p-value plots for all-cause mortality, cardiovascular mortality, and all cancer mortality appear bilinear, hence ambiguous. The p-value plots for breast cancer incidence and colorectal cancer incidence appear as 45º lines, suggesting a likelihood of no effect.

The p-value plot for colorectal cancer incidence (bottom left-hand side) is unusual, with the seven largest p-values on a roughly 45º line, two below the 0.05 threshold, and one extremely small p-value (6.2×10^{-5}). Researchers usually take a p-value less than 0.001 as very strong evidence of a real effect (Boos & Stefanski, 2011). Others suggest that small p-values indicate failures of research integrity (Al-Marzouki et al., 2005; Roberts et al., 2007). If the small p-values indicates a real effect, then p-values larger than 0.05 should be rare.

The p-value plot for Type 2 diabetes incidence (bottom right-hand side) has an appearance of a real effect – most of the p-values are small. However, the two smallest p-values – 4.1×10^{-9} and 1.7×10^{-7}, shown as triangles – have conflicting results. The first is for a decrease of effect and the second is for an increase of effect. Our plot might suggest some support for a real association between red or processed meat and Type 2 diabetes—but with a sensible warning of conflicting results of the two smallest p-values. Here we would note a caution about possible failures of research integrity related to the base papers with small p-values as suggested by others (Roberts et al., 2007; Redman, 2013).

Each health outcome presented in Figure 1 displays a wide range of p-value results – refer to Table 3. In the

meta-analysis of breast cancer incidence (middle right-hand side), for example, p-values ranged from < 0.005 to 1 across 19 base papers (> 2 orders of magnitude). In the meta-analysis of Type 2 diabetes incidence (bottom right-hand side), the p-values ranged from $< 5 \times 10^{-9}$ to 0.43 (> 7 orders of magnitude) – which suggests possible research integrity issues associated with small p-value results.

Table 3. Minimum and maximum p-values for six health outcomes shown in Figure 1 from Vernooij.

Health outcome	Number of p-values	Minimum p-value	Maximum p-value
All-cause mortality	25	1.6E-08	1
Cardiovascular mortality	27	6.4E-06	0.82
Overall cancer mortality	19	0.00032	0.89
Breast cancer incidence	19	0.0024	1
Colorectal cancer incidence	19	6.2×10^{-5}	0.78
Type 2 diabetes incidence	16	4.1×10^{-9}	0.43

The smallest p-value from Table 6 is 4.1×10^{-9} – a value small enough to imply certainty (Boos & Stefanski, 2011). A p value this small may register a true finding – and small p-values are more likely in studies with large sample sizes (Young, 2008). But the wide range of p-values in studies asking the exact same research question – including several studies which register results far weaker than $p < 0.05$ – suggests that alternative explanations cannot be ruled out. These explanations may include some form of QRP – ranging from bias (e.g., alteration of research findings due to factors related to study design, data acquisition, and/or analysis or reporting of results) (Ioannidis, 2008) all the way to data fraud and fabrication (Mojon-Azzi & Mojon, 2004; Eisenach, 2009; George & Buyse, 2015).

3.3 Research Questions Asked of Cohort Populations

Table 4 shows how common FFQ data is used by researchers investigating health outcomes in the Google Scholar literature for 18 health effects we selected (search performed 22 March 2021). Obesity associated with foods is a particular topic of interest with researchers. However, outcomes less commonly expected to be related to foods, e.g., reproductive outcomes and erectile dysfunction, have been investigated.

Table 5 presents the 15 cohorts and an estimate of the number of articles in Google Scholar literature for each cohort using FFQs (search performed 27 May 2021). From Table 5 we suggest that researchers overall may examine many hypotheses on a single cohort−FFQ data set and possibly without proper attention to multiple testing bias. We use the example of the Adventist Health Study-2 cohort data set from Table 5 to demonstrate the potential problem. If 653 studies were published on this cohort population data set using FFQs and each study examined approximately 20,000 hypotheses (i.e., similar to the median number of hypotheses in Table 2), 5% of 653 × 20,000 hypotheses equals 653,000 chance findings that may be mistaken for real results across these studies.

4. Discussion

Regarding red meat−disease association studies, others report that red and processed meat consumption is associated with adverse health effects (e.g., Battaglia et al., 2015; Ekmekcioglu et al., 2015). The International Agency for Research on Cancer (IARC), the cancer agency of the World Health Organization, has classified red meat as probably carcinogenic to humans and processed meat as certainly carcinogenic to humans (WHO, 2015). We have stated previously that performance of another study statistically confirming the same hypothesis or claim is a cornerstone of science. The Vernooij meta-analysis offered scientific explanations against red and processed meat–health effect claims. Our independent findings suggest that the base papers used in Vernooij, properly examined statistically for false positives and possible evidence of QRP (i.e., counting of hypotheses and p-value plots), do not support the reliability of red and processed meat–health effect claims.

Examining large numbers of hypotheses without offering all findings (now possible with supplemental material and web posting) makes it challenging to discover how many true or false-positive versus null findings might exist in a single study (or indeed multiple studies using the same cohort data set). A proposal for reporting meta-analysis of observational studies in epidemiology was provided for researchers in the *Journal of the American Medical Association* (Stroup et al., 2000). This proposal is frequently acknowledged in published literature (15,612 Google Scholar citations as of 1 May 2021). However, this proposal makes no mention of multiple testing bias in observational studies, and it offers no recommendations to control for this bias. Procedures to control multiple testing bias are well-established in literature (some examples include Westfall & Young, 1993; Benjamini & Hochberg, 1995; Schaffer, 1995).

Table 4. Google Scholar search of health effects associated with foods in Web literature

RowID	Outcome (effect) of interest	# of citations
1	obesity	42,600
2	inflammation	23,100
3	depression	18,000
4	mental health	10,900
5	all-cause mortality	10,700
6	high blood pressure	9,470
7	lung and other cancers	7,180
8	metabolic disorders	5,480
9	low birth weight	4,630
10	pneumonia	2,140
11	autism	2,080
12	suicide	1,840
13	COPD	1,800
14	ADHD	1,370
15	miscarriage	1,240
16	atopic dermatitis	938
17	reproductive outcomes	537
18	erectile dysfunction	359

Note: Performed on 22 March 2021; Google Scholar search is only an approximation as Web literature changes rapidly, small changes in search specifications can change results.

Meta-analyses may provide greater evidentiary value if they combine test statistics from base papers that use reliable data and analysis procedures and, crucially, all studies are responding to the same process (Fisher, 1950; DerSimonian & Laird, 1986). Base papers that examine many hypotheses and do not correct for multiple testing bias cannot be considered reliable data for meta-analyses. Furthermore, meta-analyses that combine test statistics from base papers that do and do not correct for this bias are not combining comparable statistics.

Bilinear p-value plots in Figure 1 suggest evidence that nutritional epidemiological meta-analyses have combined test statistics from base studies that do not use comparable methods. Alternately, the bilinear plots may register the existence of one or more powerful covariates correlated with a cause (predictor variable) in some of the studies – that, for example, cardiorespiratory fitness is confounded with dietary risk of mortality (Héroux et al., 2010). However, the existence of an unrecognized covariate would also render meta-analysis' results unreliable.

Large numbers of hypotheses examined in the 15 random base papers of Vernooij – refer to Tables 1 and 2 – make it plausible to infer that some test statistics with small p-values among the base papers may be derived from some form of QRP. The large number of articles resulting from these cohort data sets (Tables 5) supports this.

Epidemiology studies that examine many hypotheses tend to provide results of limited quality for each association due to limited exposure assessment and inadequate information on potential confounders (Savitz & Olshan, 1995). These studies are prone to seek out small but (nominally) significant risk factor–health outcome associations (i.e., those that are less than 0.05) in multiple testing environments. These practices may render research susceptible to reporting false-positives as real results, and to risk mistaking an improperly controlled covariate for a positive association. A set of base studies in a meta-analysis where possible numbers of hypotheses examined are large and whose p-values demonstrate bilinearity in a p-value plot should be regarded as questionable.

We note the following limitations of our methods: counting of the possible number of hypotheses examined is not easy as the statistical details of a base study may be presented anywhere in the article or not at all; the counting formula is only an approximation; we did not include possible interactions among the variables; the use of a p-value plot for evaluation of a meta-analytic result is relatively new; and the Google Scholar searches are only approximations of numbers of articles in Web literature as the Web literature changes rapidly and small changes in search specifications can change results.

Table 5. Cohort study names and estimate of papers in Web literature using FFQs for the 15 randomly sampled base papers of Vernooij

Citation#	Author	Year	Cohort Study Name	Papers, Cohort+FFQ
48	Chan	2013	Mr. Os and Ms Os (Hong Kong)	8
56	George	2014	WHI Women's Health Initiative Observational Study	1,520
49	Chen	2013	HEALS and 'Bangladesh'	1,080
53	Maruyama	2013	JACC Japan Collaborative Cohort	758
57	Kumagai	2014	NHI Ohsaki National Health Insurance Cohort	122
47	Akbaraly	2013	Whitehall II study	1,800
99	Tonstad	2013	Adventist Health Study-2	653
80	Lv	2017	China Kadoorie Biobank	143
59	Pastorino	2016	MRC National Survey of Health and Development	148
31	McNaughton	2009	Whitehall II study	1,800
34	Panagiotakos	2009	ATTICA Study	1,650
8	Dixon	2004	DIETSCAN (Dietary Patterns and Cancer Project)	1,080
38	Héroux	2010	ACLS (Aerobics Center Longitudinal Study)	167
65	Lacoppidan	2015	Diet, Cancer, and Health (DCH) cohort	116
92	Chang-Claude	2005	German vegetarian study	13

Note: Google Scholar search performed 17 May 2021; Citation# = Vernooij reference number; Author name = first author listed for reference; Year = publication year; Cohort Name = name of study cohort; Papers, Cohort + FFQ = # of papers in literature mentioning study cohort using a Food Frequency Questionnaire (FFQ); Google Scholar search is only an approximation as Web literature changes rapidly, small changes in search specifications can change results.

5. Findings

We independently evaluated the Vernooij meta-analysis. Specifically, we examined properties of the test statistics that were combined to derive meta-analytic statistical associations between red and processed meat and all-cause mortality, cardiovascular mortality, all cancer mortality, breast cancer incidence, colorectal cancer incidence, type 2 diabetes incidence. The possible number of hypotheses examined in 15 random base papers we evaluated was large, median = 20,736 (interquartile range = 1,728–331,776). Each p-value plot of the test statistics for selected health effects we evaluated showed either a random pattern (p-values > 0.05), or a two-component mixture with small p-values < 0.001 while other p-values appeared random. Given potentially large numbers of hypotheses examined in the base papers, questionable research practices cannot be ruled out as explanations for test statistics with small p-values. Given this evidence, we conclude that: i) our statistical examination does not support the reliability of red meat−negative health claims, and ii) the Vernooij finding – *…dietary patterns with less red and processed meat intake may result in very small reductions in adverse cardiometabolic and cancer outcomes* – is reliable.

Acknowledgments

We thank Bradley Johnston (Department of Nutrition, College of Agriculture and Life Sciences, Texas A&M University, College Station, TX) for providing us with the datasets used in our research. 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

Al-Marzouki, S., Evans, S., & Marshall, T., et al. (2005). Are these data real? Statistical methods for the detection of data fabrication in clinical trials. *British Medical Journal, 331*, 267. https://doi.org/10.1136/bmj.331.7511.267

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

Arends, B. (2020). *'Totally bizarre!'—nutritionists see red over study downplaying the serious health risks of red meat*. MarketWatch. https://www.marketwatch.com/story/nutritionists-see-red-over-study-downplaying-the-health-risks-of-red-meat-20

19-10-02

Battaglia Richi, E., Baumer, B., & Conrad, B., et al. (2015). Health risks associated with meat consumption: A review of epidemiological studies. *International Journal for Vitamin and Nutrition Research*, *85*(1−2), 70−78. http://dx.doi.org/10.1024/0300-9831/a000224

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *57*(1), 125−133. http://dx.doi.org/10.1111/j.2517-6161.1995.tb02031.x

Boeing H. (2013). Nutritional epidemiology: New perspectives for understanding the diet-disease relationship? *European Journal of Clinical Nutrition*, *67*(5), 424−429. http://dx.doi.org/0.1038/ejcn.2013.47

Boffetta, P., McLaughlin, J. K., & Vecchia, C. L., et al. (2008). False-positive results in cancer epidemiology: A plea for epistemological modesty. *Journal of the National Cancer Institute*, *100*, 988−995. http://dx.doi.org/10.1093/jnci/djn191

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

Bordewijk, E. M., Wang, R., & Aski, L. M., et al. (2020). Data integrity of 35 randomised controlled trials in women' health. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, *249*, 72−83. http://dx.doi.org/10.1016/j.ejogrb.2020.04.016

Creswell, J. (2003). *Research Design-Qualitative, Quantitative and Mixed Methods Approaches, 2nd ed*. Thousand Oaks, CA: Sage Publications.

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

Dickersin, K., & Chalmers, I. (2011). Recognizing, investigating and dealing with incomplete and biased reporting of clinical research: From Francis Bacon to the WHO. *Journal of the Royal Society of Medicine*, *104*, 532−538. http://dx.doi.org/10.1258/jrsm.2011.11k042

Egger, M., Davey Smith, G., & Altman, D. 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: BMJ Books. https://doi.org/10.1002/9780470693926

Eisenach, J. C. (2009). Data fabrication and article retraction: how not to get lost in the woods. *Anesthesiology*, *110*(5), 955−956. http://dx.doi.org/10.1097/ALN.0b013e3181a06bf9

Ekmekcioglu, C., Wallner, P., & Kundi, M., et al. (2018). Red meat, diseases, and healthy alternatives: A critical review. *Critical Reviews in Food Science and Nutrition*, *8*(2), 247−261. http://dx.doi.org/10.1080/10408398.2016.1158148

Fisher, R. A. (1950). *Statistical Methods for Research Workers, 11th ed*. pp 99−101. Edinburgh: Oliver and Boyd.

Frieden, T. R. (2017). Evidence for health decision making – beyond randomized, controlled trials. *New England Journal of Medicine*, *377*(5), 465−475. http://dx.doi.org/10.1056/NEJMra1614394

George, S. L., & Buyse, M. (2015), Data fraud in clinical trials. *Clinical Investigation (London)*, *5*(2), 161−173. http://dx.doi.org/10.4155/cli.14.116

George, S. M., Ballard-Barbash, R., & Manson, J. E., et al. (2014). Comparing indices of diet quality with chronic disease mortality risk in postmenopausal women in the Women's Health Initiative Observational Study: Evidence to inform national dietary guidance. *American Journal of Epidemiology*, *180*(6), 616−625. http://dx.doi.org/10.1093/aje/kwu173

Gotzsche, P. C. (2006). Believability of relative risks and odds ratios in abstracts: Cross sectional study. *British Medical Journal*, *333*, 231−234. http://dx.doi.org/10.1136/bmj.38895.410451.79

Guyatt, G. H., Oxman, A. D., & Vist, G. E., et al; GRADE Working Group. (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal*, *336*, 924. http://dx.doi.org/10.1136/bmj.39489.470347.AD

Héroux, M., Janssen, I., & Lam, M., et al. (2010). Dietary patterns and the risk of mortality: Impact of cardiorespiratory fitness. *International Journal of Epidemiology*, *39*(1), 197−209. http://dx.doi.org/10.1093/ije/dyp191

Hung, H. M. J., O'Neill, R. T., & Bauer, P., et al. (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. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640−648. http://dx.doi.org/10.1097/EDE.0b013e31818131e7

Kunert, R. (2016). Internal conceptual replications do not increase independent replication success. *Psychonomic Bulletin & Review*, *23*(5), 1631−1638. http://dx.doi.org/10.3758/s13423-016-1030-9

Malik, V. S., Popkin, B.,& Bray, G., et al. (2010). Sugar-sweetened beverages and risk of metabolic syndrome and type 2 diabetes. *Diabetes Care*, *33*, 2477−2483. http://dx.doi.org/10.2337/dc10-1079

Moher, D., Liberati, A., & Tetzlaff, J., et al. (2009). The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*, e1000097. http://dx.doi.org/10.1371/journal.pmed.1000097

Mojon-Azzi, S. M., & Mojon, D. S. (2004). Scientific misconduct: from salami slicing to data fabrication. *Ophthalmic Research*, *36*(1), 1−3. http://dx.doi.org/10.1159/000076104

Monaco, K. (2019). *Is everything we know about meat consumption wrong?* Medpage Today. https://www.medpagetoday.com/primarycare/dietnutrition/82492

Moonesinghe, R., Khoury, M. J., & Janssens, A. C. J. W. (2007). Most published research findings are false−But a little replication goes a long way. *PLoS Medicine*, *4*(2), e28. http://dx.doi.org/10.1371/journal.pmed.0040028

Murad, M. H., Asi, N., & Alsawas, M., et al. (2016). New evidence pyramid. *BMJ Evidence-Based Medicine*, *21*(4), 125−127. http://dx.doi.org/10.1136/ebmed-2016-110401

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

Panagiotakos, D., Pitsavos, C., & Chrysohoou, C., et al. (2009). Dietary patterns and 5-year incidence of cardiovascular disease: a multivariate analysis of the ATTICA study. *Nutrition, Metabolism & Cardiovascular Diseases*, *19*(4), 253−263. http://dx.doi.org/10.1016/j.numecd.2008.06.005

Peace, K. E., Yin, J. J., & Rochani, H., et al. (2018). A serious flaw in nutrition epidemiology: A meta-analysis study. *International Journal of Biostatistics*, *14*(2), pp. 20180079. http://dx.doi.org/10.1515/ijb-2018-0079

Prentice, R. L., & Huang, Y. (2018). Nutritional epidemiology methods and related statistical challenges and opportunities. Statistical *Theory and Related Fields*, *2*(1), 2−10. http://dx.doi.org/10.1080/24754269.2018.1466098

Potischman, N., & Weed, D. L. (1999). Causal criteria in nutritional epidemiology. *American Journal of Clinical Nutrition*, *69*(6), 1309S−1314S. http://dx.doi.org/10.1093/ajcn/69.6.1309S

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

Redman, B. K. (2013). *Research Misconduct Policy in Biomedicine; Beyond the Bad Apple Approach*. Cambridge, MA: The MIT Press.

Roberts, I., Smith, R., & Evans, S. (2007). Doubts over head injury studies. *British Medical Journal*, *334*, 392. http://dx.doi.org/10.1136/bmj.39118.480023.BE

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

Satija, A., Yu, E., Willett, W. C., et al. (2015). Understanding nutritional epidemiology and its role in policy. *Advances in Nutrition*, *6*(1), 5−18. http://dx.doi.org/10.3945/an.114.007492

Savitz, D. A., & Olshan, A. F. (1995). Multiple comparisons and related issues in the interpretation of epidemiologic data. *American Journal of Epidemiology*, *142*, 904−908. http://dx.doi.org/10.1093/oxfordjournals.aje.a117737

Schaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*, 561−584. http://dx.doi.org/10.1146/annurev.ps.46.020195.003021

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

Stroup, D. F., Berlin, J. A., & Morton, S. C., et al. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. *Journal of the American Medical Association*, *283*(15), 2008−2012.

http://dx.doi.org/10.1001/jama.283.15.2008

Vernooij, R. W. M., Zeraatkar, D., & Han, M. A., et al. (2019). Patterns of red and processed meat consumption and risk for cardiometabolic and cancer outcomes A systematic review and meta-analysis of cohort studies. *Annals of Internal Medicine*, *171*, 732−741. http://dx.doi.org/10.7326/M19-1583

Ware, J. J., & Munafò, M. R. (2015). Significance chasing in research practice: causes, consequences and possible solutions. *Addiction*, *110*(1), 4−8. http://dx.doi.org/10.1111/add.12673

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

Willett, W. C., Sampson, L., & Stampfer, M. J., et al. (1985). Reproducibility and validity of a semiquantitative food frequency questionnaire. *American Journal of Epidemiology*, *122*, 51−65. http://dx.doi.org/10.1093/oxfordjournals.aje.a114086

World Health Organization (WHO). (2015). *IARC Monographs evaluate consumption of red meat and processed meat. Press release No. 240*. Lyon, France: International Agency for Research on Cancer (IARC), World Health Organization. https://www.iarc.who.int/wp-content/uploads/2018/07/pr240_E.pdf

Young, S. S. (2008). *Statistical analyses and interpretation of complex studies*. Medscape. https://www.medscape.org/viewarticle/571523

Young, S.S., Cheng, K.-C., & Chen, J. H., et al. (2022). Reliability of a meta-analysis of air quality−asthma cohort studies. *International Journal of Statistics and Probability*, *11*(2), 61−76. https://doi.org/10.5539/ijspv11n2p61

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., & Randall, D. (2021). *Shifting Sands, Unsound Science and Unsafe Regulation Report 1. Keeping Count of Government Science: P-Value Plotting, P-Hacking, and PM2.5 Regulation*. New York, NY: National Association of Scholars. https://www.nas.org/reports/shifting-sands-report-i