Downloaded from http://rspb.royalsocietypublishing.org/ on October 23, 2018

PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2009) **276**, 1211–1212
doi:10.1098/rspb.2008.1405
Published online 13 January 2009

Comment

Cereal-induced gender selection? Most likely a multiple testing false positive

The recent paper by Mathews *et al.* (2008) with a provocative title 'You are what your mother eats' generated a lot of attention in the press and over 50 000 Google hits putting forth the genetically implausible claim that women who eat breakfast cereal are more likely to have a boy child. Their result is easily explained as chance. We will not go into other methodological issues such as recall bias and measurement errors, difficulty in measuring cumulative exposures in nutritional data, unmeasured confounders, variable categorization, statistical power and study design, as Pocock *et al.* (2004) recently reviewed the sad state of observational studies and Ioannidis (2005) reports that 80 per cent of observational studies fail to replicate or the initial effects are much smaller on retest. An implausible claim should strongly overcome chance as an explanation even to be considered. We focus on chance as the cause of their finding.

It has been long well-known, Cournot (1843), that multiple testing can easily lead to false discoveries when multiple hypothesis testing or comparisons are not adequately taken into account. Cournot commented, 'One could distinguish first of all legitimate births from those occurring out of wedlock, … one can also classify births according to birth order, according to the age, profession, wealth or religion of the parents…' Cournot goes on to point out that as one increases the number of such 'cuts' (of the material into two or more categories) it becomes more and more likely that by pure chance for at least one pair of opposing categories the observed difference will be significant. Based on a careful reading of the paper by Mathews *et al.* (2008), a counting of the questions under consideration and an analysis that better takes multiple testing into account, we strongly believe that the main finding in this paper to be a false discovery/type I error. Hundreds of comparisons were conducted; there also seems to be hidden multiple testing as many additional tests were computed and reported in other papers.

Specifically, the authors state in the abstract of their paper 'Fifty six per cent of women in the highest third of preconceptional energy intake bore boys, compared with 45 per cent in the lowest third' and assert that this result is statistically significant. They go on to the point of breakfast cereal consumption for the prediction of infant gender. The authors provided the dataset and we conducted our own analysis looking at the individual food items for time periods one and two, $132 \times 2 = 264$ statistical tests. (Note that there are actually three time periods and only 132 food items were actually present in the supplied dataset, so there are nominally $132 \times 3 = 396$ questions at issue.) There was a third time period, but the authors did not present data from this period (table 2). In our first analysis, we computed 264 *t*-tests and plotted the resulting ordered *p*-values versus the integers giving a *p*-value plot, Schweder & Spjøtvoll (1982); figure 1. Some explanation: suppose we statistically test 10 questions where nothing is going on. By chance alone we expect the smallest *p*-value to be rather small. We actually expect the *p*-values to be nicely spread out uniformly over the interval 0–1. Except for sampling variability, we expect that the ordered *p*-values plotted against the integers, 1, 2, …, 10, to line up along a 45-degree line. With this dataset, we have 264 *p*-values and the plot of the ordered *p*-values against the integers, 1, 2, …, 264 is essentially linear. This plot implies that the small observed *p*-values, indeed all of the *p*-values, are simply the result of chance and not due to any effect of the food items.

In our second analysis, we used simulation to compute multiplicity-adjusted *p*-values. Explanation of the computation of adjusted *p*-values: we would wish to know if the smallest observed *p*-value could have arisen by chance. We take the outcome for each mother, 0/1 for girl/boy, and permute the values assigning the gender of the child at random to the mother. We compute *p*-values for all the food items and the smallest *p*-value in the permuted dataset is clearly a chance value. We do this permutation thousands of times and get the distribution of the smallest *p*-value. We note where the observed smallest *p*-value falls in this distribution. Within sampling error that can be made arbitrarily small, the adjusted *p*-value is the correct probability of seeing a *p*-value as small when observed, Westfall & Young (1993). The method takes into account multiple testing, the correlation structure among the variables and the distributional characteristics of the variables. For the preconception time period, the unadjusted *p*-value for breakfast cereal 0.0034 has a multiple testing adjusted *p*-value of 0.2813. This adjusted *p*-value is interpreted as follows: one would expect to see a *p*-value as small as 0.0034 approximately 28 per cent of the time when nothing is going on. So looking at both time periods using the *p*-value plot and at the individual food items in the preconception period using multiple-testing adjusted *p*-values, the claimed effects are readily explainable by chance. In addition, the motivating small *p*-values in table 2, are also explainable by chance. The authors report an unadjusted *p*-value of 0.029 for total energy. Among 54 tests, a *p*-value of 0.029 is not unusual, so total energy is not statistically significant. Interestingly, sodium gave the smallest *p*-value in table 2, an unadjusted *p*-value of 0.003 (which the authors dismiss); this *p*-value is also not statistically significant when adjusted for multiple testing.

The accompanying reply can be viewed on page 1213 or at http://dx.doi.org/doi:10.1098/rspb.2008.1781.

Received 29 September 2008
Accepted 24 October 2008

This journal is © 2009 The Royal Society

Downloaded from http://rspb.royalsocietypublishing.org/ on October 23, 2018

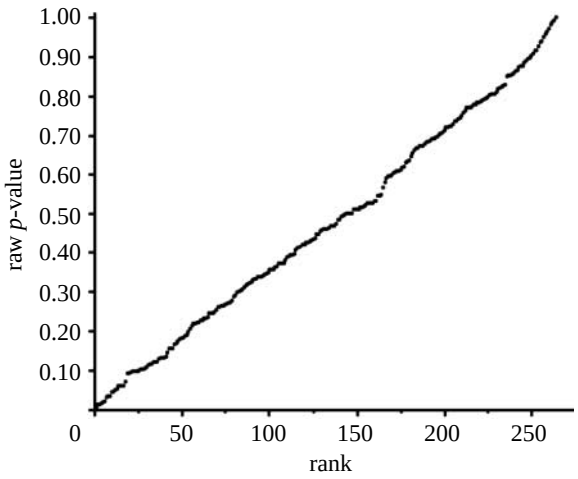

Figure 1. The *p*-value plot of 262 *p*-values.

So the motivating test from table 2, total energy, is not significant and the individual food item, breakfast cereal, is also not significantly associated with the outcome, the gender of a baby. The authors' claim that a principal component analysis used as a 'gate keeper' protects them from making a false positive claim. The resampling-based analysis corrects perfectly the correlation structure and distributional characteristics and shows that the *p*-values claimed as significant by the authors are easily the result of chance. In addition, the *p*-value plot supports the conclusion that all the *p*-values are explainable by chance.

Lest the reader thinks multiple testing is not important, we mention two historic examples. In the 1970s, many diseases were reported to be associated with an human leukocyte antigen (HLA) allele (schizophrenia, hypertension… you name it!). Researchers did case-control studies with 40 antigens, so there was a very high chance, approximately 87 per cent, of at least one significant result in any study. Any result was reported without mention of the fact that it was the most significant of 40 tests (R. C. Elston 2008, personal communication). Westfall (1985) provided a solution to multiple testing for HLA analysis. Another example is the reported association between reserpine (then a popular antihypertensive) and breast cancer. Shapiro (2004) gives the history. His team published initial results linking reserpine and breast cancer which were extensively covered by the media with a huge impact on the research community at the time. When the results did not replicate, he came to the conclusion that the initial findings were chance due to thousands of comparisons involving hundreds of outcomes and hundreds of different drugs under consideration. He hopes that we learn from his mistake. Given that the prevailing observational study paradigm is not to correct the multiple testing, Shapiro is indeed a brave admit and speaker against what he considers a bad practice.

With Mathews *et al.* (2008) a rather complex dataset is examined with massive statistical testing and a thread is found to make a case for causality. Arguments about biological plausibility should be viewed with some scepticism since the human imagination seems capable of developing a rationale for most findings, however, unanticipated, so called retrospective rationalization (Ware 2003). Again, note the authors' 'Sodium intake is difficult to measure' rationalization for the dismissal of the smaller *p*-value for sodium.

Conventional genetics, X/Y from the male being the determinate of gender in humans, and the fact that so many tests were computed point to the authors' claims and being best explained as chance.

This paper comes across as well-intended, but it is hard to believe that women can increase the likelihood of having a baby-boy instead of a baby-girl by eating more bananas, cereal or salt. Nominal statistical significance, unadjusted for multiple testing, is often used to lend plausibility to a research finding; with an arguably implausible result, it is essential that multiple testing to be taken into account with transparent methods for claims to have any level of credibility.

Lastly, it is difficult to undo a false positive claim; it takes time to respond to a literature claim and most often the original data are not made available, Kaiser (2008). Editors and referees should consider multiple testing as a possible explanation for improbable claims. Researchers reporting on observational studies should work to a higher standard: Is this claim probable to replicate? Since observational studies are subject to many problems and most do not adjust for multiple testing, readers would be well advised to ignore claims from observational studies until replicated.

S. Stanley Young[1,*], Heejung Bang[2] and Kutluk Oktay[3]
[1]*National Institute of Statistical Sciences, Research Triangle Park, NC 27709, USA*
[2]*Cornell University, New York, NY 10021 USA*
[3]*New York Medical College, Valhalla, NY 10595, USA*
*Author for correspondence (young@niss.org).

REFERENCES

Cournot, A. A. 1843 *Exposition de la Theorie des Chance des Probability*. Paris, France: Librairie de L'Hachette.

Ioannidis, J. P. A. 2005 Contradicted and initially stronger effects in highly cited clinical research. *JAMA* **294**, 218–228. (doi:10.1001/jama.294.2.218)

Kaiser, J. 2008 Making clinical data widely available. *Science* **322**, 217–218. (doi:10.1126/science.322.5899.217)

Mathews, F., Johnson, P. J. & Neil, A. 2008 You are what your mother eats: evidence for maternal preconception diet influencing foetal sex in humans. *Proc. R. Soc. B* **275**, 1661–1668. (doi:10.1098/rspb.2008.0105)

Pocock, S. J., Collier, T. J. & Dandreo, K. J. 2004 Issues in the reporting of epidemiological studies: a survey of recent practice. *BMJ* **329**, 883–888. (doi:10.1136/bmj.38250.571088.55)

Schweder, T. & Spjøtvoll, E. 1982 Plots of *p*-values to evaluate many tests simultaneously. *Biometrika* **69**, 493–502. (doi:10.2307/2335984)

Shapiro, S. 2004 Looking to the 21st century: have we learned from our mistakes, or are we doomed to compound them? *Pharmacoepidemiol. Drug Saf.* **13**, 257–265. (doi:10.1002/pds.903)

Ware, J. H. 2003 The national emphysema treatment trial: how strong is the evidence? *NEJM* **348**, 2055–2056. (doi:10.1056/NEJMp030068)

Westfall, P. 1985 Simultaneous small-sample multivariate Bernoulli confidence intervals. *Biometics* **41**, 1001–1013. (doi:10.2307/2530971)

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