Deming, data and observational studies

A process out of control and needing fixing

"Any claim coming from an observational study is most likely to be wrong." Startling, but true. Coffee causes pancreatic cancer. Type A personality causes heart attacks. Trans-fat is a killer. Women who eat breakfast cereal give birth to more boys. All these claims come from observational studies; yet when the studies are carefully examined, the claimed links appear to be incorrect. What is going wrong? Some have suggested that the scientific method is failing, that nature itself is playing tricks on us. But it is our way of studying nature that is broken and that urgently needs mending, say **S. Stanley Young** and **Alan Karr**; and they propose a strategy to fix it.

Science works by experiments that can be repeated; when they are repeated, they must give the same answer. If an experiment does not replicate, something has gone wrong. In a large branch of science the experiments are observational studies: we look at people who eat certain foods, or take certain drugs, or live certain lifestyles, and we seem to find that they suffer more from certain diseases or are cured of those diseases, or – as with women who eat more breakfast cereal – that more of their children are boys. The more startling the claim, the better. These results are published in peer-reviewed journals, and frequently make news headlines as well. They seem solid. They are based on observation, on scientific method, and on statistics. But something is going wrong. There is now enough evidence to say what many have long thought: that any claim coming from an observational study is most likely to be wrong – wrong in the sense that it will not replicate if tested rigorously.

As long ago as 1988[1,2] it was noted that there were contradicted results for case–control studies in 56 different topic areas, of which

Table 1. We have found 12 papers in which claims coming from observational studies were tested in randomised clinical trials. Many of the trials are quite large. In most of the observational studies multiple claims were tested, often in factorial designs, e.g. vitamin D and calcium individually and together along with a placebo group. Note that none of the claims replicated in the direction claimed in the observational studies and that there was statistical significance in the opposite direction five times

ID no.	*Pos.*	*Neg.*	*No. of claims*	*Treatment(s)*	*Reference*
1	0	1	3	Vit E, beta-carotene	*NEJM* 1994; **330**: 1029–1035
2	0	3	4	Hormone Replacement Ther.	*JAMA* 2003; **289**: 2651–2662, 2663–2672, 2673–2684
3	0	1	2	Vit E, beta-carotene	*JNCI* 2005; **97**: 481–488
4	0	0	3	Vit E	*JAMA* 2005; **293**: 1338–1347
5	0	0	3	Low Fat	*JAMA*. 2006; **295**: 655–666
6	0	0	3	Vit D, Calcium	*NEJM* 2006; **354**: 669–683
7	0	0	2	Folic acid, Vit B6, B12	*NEJM* 2006; **354**: 2764–2772
8	0	0	2	Low Fat	*JAMA* 2007; **298**: 289–298
9	0	0	12	Vit C, Vit E, beta-carotene	*Arch Intern Med* 2007; **167**: 1610–1618
10	0	0	12	Vit C, Vit E	*JAMA* 2008; **300**: 2123–2133
11	0	0	3	Vit E, Selenium	*JAMA* 2009; **301**: 39–51
12	0	0	3	HRT + Vitamins	*JAMA* 2002; **288**: 2431–2440
Totals	0	5	52		

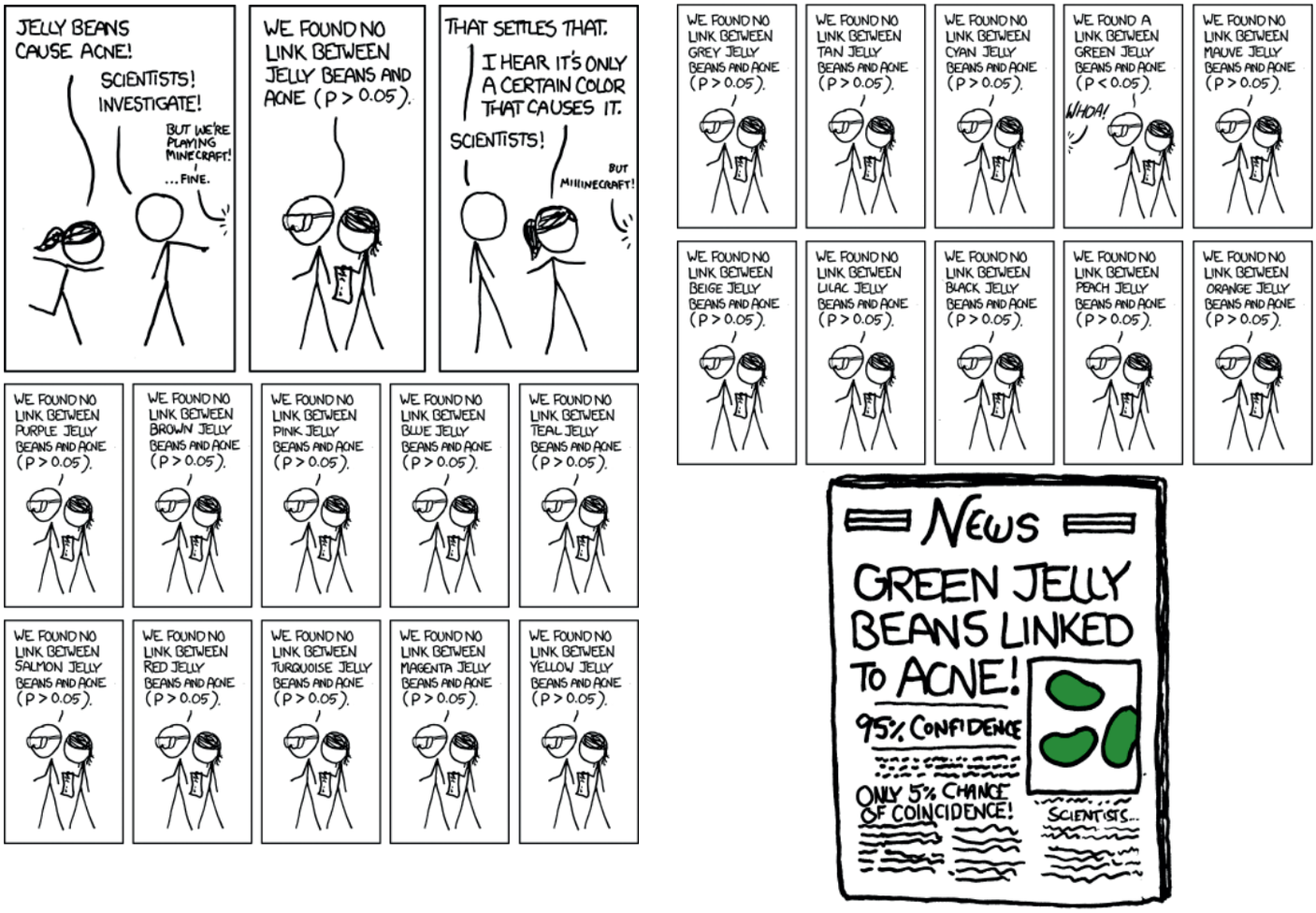


Figure 1. There is no overall effect of jelly beans on acne. Bummer. How about subgroups? Often subgroups are explored without alerting the reader to the number of questions at issue. Courtesy xkcd, http://xkcd.com/882/

cancer and things that cause it or cure it were by far the most frequent. An average of 2.4 studies supported each association – and an average of 2.3 studies did not support it. For example, three studies supported an association between the anti-depressant drug reserpine and breast cancer, and eight did not. It was asserted[2] that "much of the disagreement may occur because a set of rigorous scientific principles has not yet been accepted to guide the design or interpretation of case-control research". Problems extend to essentially all observational studies. Little progress has been made to adopt rigorous scientific principles. Some journal article titles give a flavour of the sentiments: "Epidemiology faces its limits", "Is it time to call it a day?", "Have we learned from our mistakes, or are we doomed to compound them?". In the popular press, an article by Jonah Lehrer in the *New Yorker*[3] bore the subheading "Is there something wrong with the scientific method?" and seemed to imply that replicability was no longer occurring; it concluded with the phrase: "When the experiments are done, we still have to choose what to believe." No. In the Lehrer example the motivating finding was wrong and therefore should not be expected to replicate.

It may not be appreciated how often observational claims fail to replicate. In a small sample in 2005[4], of 49 claims coming from highly cited studies, 14 either failed to replicate entirely or the magnitude of the claimed effect was greatly reduced (a regression to the mean). Six of these 49 studies were observational studies, and in these six, in effect, randomly chosen observational studies, five failed to replicate. This last is an 83% failure rate. In an ideal world in which well-studied questions are addressed and statistical issues are accounted for properly, few statistically significant claims are false positives. Reality for observational studies is quite different.

We ourselves carried out an informal but comprehensive accounting of 12 randomised clinical trials that tested observational claims – see Table 1. The 12 clinical trials tested 52 observational claims. They all confirmed no claims in the direction of the observational claims. We repeat that figure: 0 out of 52. To put it another way, 100% of the observational claims failed to replicate. In fact, five claims (9.6%) are statistically significant in the clinical trials *in the opposite direction* to the observational claim. To us, a false discovery rate of over 80% is potent evidence that the observational study process is not in control. The problem, which has been recognised at least since 1988, is systemic.

The cause of it all

The cause is elusive and can be considered both technically and operationally. Individual researchers, the workers, respond rationally to incentives by publishing papers in peer-reviewed journals and securing funding for their research. The quality of their papers is judged by funding agencies and journal editors, the important managers of the observational study production system. We can turn here to statistician W.

Box 1. Amplification of W. Edwards Deming's thinking

It is worth contrasting control of an observational study with that of a production process. When Deming first looked at manufacturing, it was common to inspect only the final product, be it a screw or a car, to maintain product quality. There was little or no systematic feedback from problems with the final product to places in the process where these defects occurred. This inspection of the final product works, but it is frightfully expensive. Deming's insight was to control each step of the process where errors occur so that the final frequency of bad product is greatly reduced. Now, world-wide, industrial production is *process control*. Control the steps of the process and the final product will largely take care of itself. Consider the production of an observational study: Workers – that is, researchers – do data collection, data cleaning, statistical analysis, interpretation, writing a report/paper. It is a craft with essentially no managerial control at each step of the process. In contrast, management dictates control at multiple steps in the manufacture of computer chips, to name only one process control example. But journal editors and referees inspect only the final product of the observational study production process and they release a lot of bad product. The consumer is left to sort it all out. No amount of educating the consumer will fix the process. No amount of teaching – or of blaming – the worker will materially change the group behaviour. Deming's insight was to admonish management to redesign an out-of-control process.

Edwards Deming[5], the most visionary innovator ever on quality control and the man who transformed first Japanese car manufacturing then manufacturing quality control worldwide (see Box 1). Deming said: "The worker is not the problem. The problem is at the top! Management!" To Deming, blaming the workers – individual researchers – is as incorrect as it is useless. Bringing the system under control is the responsibility of those managing it.

What is needed to fix the system? Among Deming's famous "Fourteen Points for Management", the third is most directly relevant: *cease dependence on inspection to achieve quality*. Every successful company today relies on control of the process; they do not wait until the end of the process and then throw away bad product. That would be product control, not process control. It is wasteful to make something, then inspect and throw away the bad product. Instead, every step of the process is monitored and controlled, so that bad product is not made. The "observational studies industry" must build a good product; journal editors cannot inspect bad product out at the publication stage, let alone the replication stage. If the processes are controlled by management, the products can be sound studies. Control of the processes is feasible, and requires attention to the incentives, publications and grants. First we examine three of the main technical difficulties with observational studies: Multiple testing, bias, and multiple modelling.

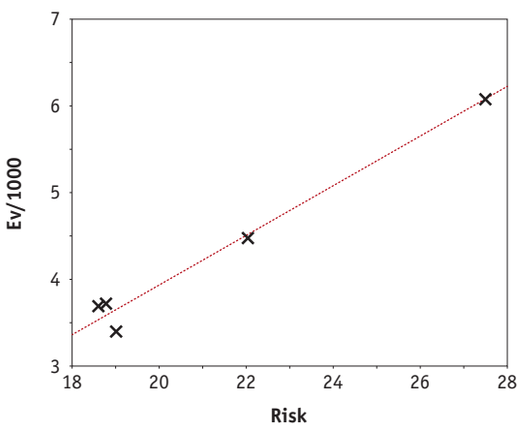
Multiple testing

False positives do occur, even in an ideal world. When many questions are asked of the same data, some of those questions will by chance come up positive. Producing at least one false positive becomes a near certainty unless the data analysis accounts for the multiple questions. Figure 1, from the excellent website `xkcd.com`, brilliantly explains the basic problem. The "females eating cereal leads to more boy babies" claim translated the cartoon example into real life. The claim appeared in the *Proceedings of the Royal Society, Series B*. It makes essentially no biological sense, as for humans the Y chromosome controls gender and comes from the male parent. The data set consisted of the gender of children of 740 mothers along with the results of a food questionnaire, not of breakfast cereal alone but of 133 different food items – compared to only 20 colours of jelly beans. Breakfast cereal during the second time period at issue was one of the few foods of the 133 to give a positive. We reanalysed the data[6], with 262 *t*-tests, and concluded that the result was easily explained as pure chance.

For those who want more than cartoons, a simple web simulation[7] is convincing that multiple testing needs to be controlled. Although many workers who are thought leaders of researchers doing observational studies argue against any correction of the analysis for multiple testing[8], managers can require that authors deal with multiple testing.

Bias

Whereas multiple testing is random error, bias is systematic error. To illustrate it, consider channelling, where doctors steer certain patients to particular treatments. For example, doctors directed HIV patients at high cardiovascular risk to a particular HIV treatment, abacavir, and lower-risk patients to other drugs, preventing a simple assessment of abacavir compared to other treatments. An analysis that did not correct for this bias unfairly penalised the abacavir, since its patients were more high-risk so more of them had heart attacks (Figure 2). Another problem is that covariate adjustment is widely used, but is vulnerable to manipulation and is well known to give unreliable results when the treatment groups are not comparable; see "Multiple modelling" below. Missing factors, unmeasured confounders, and loss to follow-up can also lead to bias. For example, in a study published in *Pediatrics*[9], offspring IQ was the issue, yet IQ of the fathers was not measured and of the 505 children starting the study, 256 (50.7%) were lost to follow-up. By selecting papers with a significant *p*-value, negative studies are selected against – which is publication bias (see Box 2).



	Drug	% high-risk patients	MI events per 1000 patient years
1	zidovudine	19.0	3.43
2	stavudine	18.8	3.72
3	lamivudine	18.6	3.68
4	didanosine	22.1	4.47
5	abacavir	27.5	6.13

Figure 2. Events per thousand patient-years are plotted against estimated risk of a heart attack. Risky patients were channelled to the HIV drug ABC, abacavir, and those patients had more heart attacks, as shown by the uppermost point on the graph. Risk-adjusted, all the drugs appear to be of equal risk. Source: *Lancet* **371**, 1417 ff.

Box 2. Publication bias

There is general recognition that a paper has a much better chance of acceptance if something new is found. This means that, for publication, the claim in the paper has to be based on a *p*-value less than 0.05. From Deming's point of view[5], this is quality by inspection. The journals are placing heavy reliance on a statistical test rather than examination of the methods and steps that lead to a conclusion. As to having a *p*-value less than 0.05, some might be tempted to game the system[10] through multiple testing, multiple modelling or unfair treatment of bias, or some combination of the three that leads to a small *p*-value. Researchers can be quite creative in devising a plausible story to fit the statistical finding.

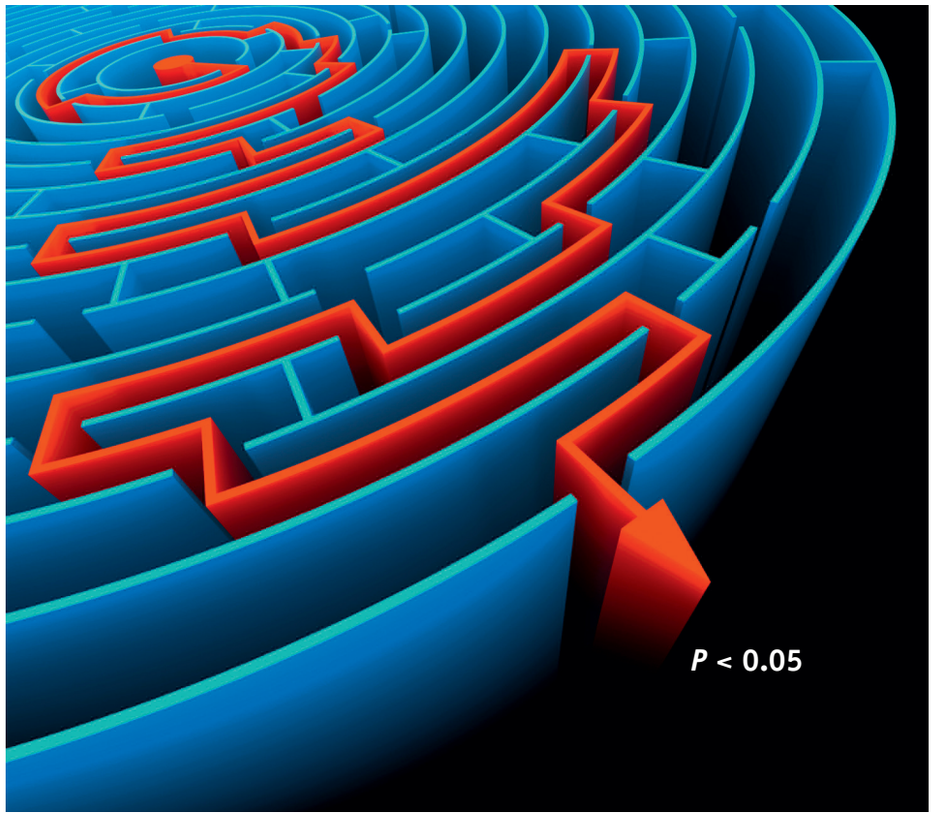


Figure 3. The path through a complex process can appear quite simple once the path is defined. Which terms are included in a multiple linear regression model? Each turn in a maze is analogous to including or not a specific term in the evolving linear model. By keeping an eye on the p-value on the term selected to be at issue, one can work towards a suitably small p-value. © ktsdesign – Fotolia

Multiple modelling

This problem is akin to – but less well recognised and more poorly understood than – multiple testing. For example, consider the use of linear regression to adjust the risk levels of two treatments to the same background level of risk. There can be many covariates, and each set of covariates can be in or out of the model. With ten covariates, there are over 1000 possible models. Consider a maze as a metaphor for modelling (Figure 3). The red line traces the correct path out of the maze. The path through the maze looks simple, once it is known. Returning to a linear regression model, terms can be put into and taken out of a regression model. Once you get a *p*-value smaller than 0.05, the model can be frozen and the model selection justified after the fact. It is easy to justify each turn.

The combination of multiple testing and multiple modelling can lead to a very large search space, as the example of bisphenol A in Box 3 shows. Such large search spaces can give small, false positive *p*-values somewhere within them. Unfortunately, authors and consumers are often like a deer caught in the headlights and take a small *p*-value as indicating a real effect.

How can it be fixed? A new, combined strategy

It should be clear by now that more than small-scale remedies are needed. The entire system of observational studies and the claims that are made from them is no longer functional, nor is it fit for purpose. What can be done to fix this broken system? There are no principled ways in the literature for dealing with model selection, so we propose a new, composite strategy. Following Deming, it is based not upon the workers – the researchers – but on the production system managers – the funding agencies and the editors of the journals where the claims are reported.

We propose a multi-step strategy to help bring observational studies under control (see Table 2). The main technical idea is to split the data into two data sets, a modelling data set and a holdout data set. The main operational idea is to require the journal to accept or reject the paper based on an analysis of the modelling data set without knowing the results of applying the methods used for the modelling set on the holdout set *and* to publish an addendum to the paper giving the results of the analysis of the holdout set. We now cover the steps, one by one.

1 The data collection and clean-up should be done by a group separate from the analysis group. There can be a temptation on the part of the analyst to do some exploratory data analysis during the data clean up. Exploratory analysis could lead to model selection bias.

2 The data cleaning team creates a modelling data set and a holdout set and gives the modelling data set, less the item to be predicted, to the analyst for examination.

Table 2. Steps 0–7 can be used to help bring the observational study process into control. Currently researchers analysing observational data sets are under no effective oversight

Step	*Process / Action*
0	Data are made publicly available
1	Data cleaning and analysis separate
2	Split sample: A, modelling; and B, holdout (testing)
3	Analysis plan is written, based on modelling data only
4	Written protocol, based on viewing predictor variables of A
5	Analysis of A only data set
6	Journal accepts paper based on A only
7	Analysis of B data set gives Addendum

3 The statistical analysis plan is written based on access to all the modelling data except the response(s) to be predicted[12].
4 The analyst writes down and files the statistical protocol. The point is that the analysis should not be guided by looking at the results of exploratory analysis. It is too easy to move predictors into and out of an evolving statistical models. Reconsider the maze (Figure 3). Given flexibility, the analyst can move the answer around. Such flexibility must be prevented.
5 The analysis is done and the paper written (see Box 2).
6 The journal agrees to accept or reject the paper without knowing the results of the analysis of the holdout data set.
7 Once that analysis is done, an addendum will be added to the paper using the specified analysis on the holdout set.

A hold-out set of data can be tested against claims; if the test fails, both author and journal stand to be embarrassed

The holdout set is the key. Both the author and the journal know there is a sword of Damocles over their heads. Both stand to be embarrassed if the holdout set does not support the original claims of the author. Both the author and the journal are at present living in a largely risk-free environment. False results may never be overturned. The claim that "Type A personality causes heart attacks" still lives and took decades to be declared invalid. Most who took the claim at face value to be true never got the word that it is not true. The myth still lives. The protocol we suggest would have scotched it at birth.

Before our steps 1–7 begin, there is another step to be made. Step 0, making data available, provides additional oversight. Note that the split-sample strategy can control multiple testing and multiple modelling, but not bias. Bias can be controlled by setting a threshold of effect, say for risk ratio a value of 3 to 4[13], of effect to be considered actionable evidence of cause and effect.

Box 3. Bisphenol A

The US Center for Disease Control assayed the urine of around 1000 people for 275 chemicals, one of which was bisphenol A (BPA). One resulting claim was that BPA is associated with cardiovascular diagnoses, diabetes, and abnormal liver enzyme concentrations. BPA is a chemical much in the news and under attack from people fearful of chemicals. The people who had their urine assayed for chemicals also gave a self-reported health status for 32 medical outcomes. For each person, ten demographic variables (such as ethnicity, education, and income) were also collected. There are 275 × 32 = 8800 potential endpoints for analysis. Using simple linear regression for covariate adjustment, there are approximately 1000 potential models, including or not including each demographic variable. Altogether the search space is about 9 million models and endpoints[11]. The authors remain convinced that their claim is valid.

Deer in Headlights. A deer caught in the headlights will freeze, much like an author or reader seeing a p-value < 0.05, and think there must be a real effect. Authors can exploit this phenomenon intentionally or fool both themselves and the reader. Illustration: Tom Boulton

What can be done?

Note that workers have known of problems since at least 1988 and have instituted none of the steps 0–7 in Table 2. Asking authors voluntarily to provide protocol, data and analysis code has been very largely ineffective. There is a real limit to what an individual can do to improve the situation, as most of us are consumers. Individuals can write letters to the editor saying that without access to data the research is largely "trust me" science. The incentives need to be changed and that can only come from the managers of the process. Managers cannot carefully examine each published claim, but funding agencies and editors can require "reproducible research". Reproducible research is research where the study protocol, the electronic data set used for the paper, and the analysis code are all publicly available. Managers can also require split-sample analysis strategies and other methods to protect against false positives. At present, researchers – and, just as important, the public at large - are being deceived, and are being deceived in the name of science. This should not be allowed to continue.

References

1. Mayes, L. C., Horwitz, R. I. and Feinstein, A. R. (1988) A collection of 56 topics with contradictory results in case-control research. *International Journal of Epidemiology*, **17**, 680–685.
2. Feinstein, A. R. (1988) Scientific standards in epidemiologic studies of the menace of daily life. *Science*, **242**, 1257–1263.
3. Lehrer, J. (2010) The truth wears off. *New Yorker*, December 13th, p. 52.
4. Ioannidis, J. P. A. (2005) Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association,* **294**, 218–228.
5. Wikipedia (2011) W. Edwards Deming. http://en.wikipedia.org/wiki/W._Edwards_Deming
6. Young, S. S., Bang, H. and Oktay, K. (2009) Cereal-induced gender selection? Most likely a multiple testing false positive. *Proceedings of the Royal Society, Series B*, **654**, 1211–1212.
7. Dallal, J. (2011) There must be something buried in here somewhere. http://www.jerrydallal.com/LHSP/multtest.htm
8. Rothman, K. J. (1990) No adjustments are needed for multiple comparisons. *Epidemiology*, **1**, 43–46.
9. Perera, F. P., Li, Z., Whyatt, R., Hoepner, L., Wang, S., Camann, D. and Rauh, V. (2009) Prenatal airborne polycyclic aromatic hydrocarbon exposure and child IQ at age 5 years. *Pediatrics*, **124**, e195–e202
10. Glaeser, E. L. (2006) Researcher incentives and empirical methods. Harvard Institute of Economic Research Discussion Paper No. 2122. http://ssrn.com/abstract=934557
11. Young, S. S. and Yu, M. (2009) To the Editor. *Journal of the American Medical Association*, **301**, 720–721.
12. Rubin, D. B. (2007) The design *versus* the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, **26**, 20–36.
13. Temple, R. (1999) Meta-analysis and epidemiologic studies in drug development and postmarketing surveillance. *Journal of the American Medical Association*, **281**, 841–844.

S. Stanley Young is the Assistant Director of Bioinformatics and Alan Karr is the Director at the National Institute of Statistical Sciences.