

Prelude

From WITCHES, FLOODS, AND WONDER DRUGS: HISTORICAL PERSPECTIVES ON RISK MANAGEMENT, by William C. Clark. "For several centuries spanning the Renaissance and Reformation, societal risk assessment meant witch hunting. Contemporary accounts record wheat inexplicably rotting in the fields, sheep dying of unknown causes, vineyards smitten with unseasonable frost, human disease and impotence on the rise. In other words, a litany of life's sorrows not very different from those which concern us today.

The institutionalized expertise of that earlier time resided with the Church. Then, as now, the experts were called upon to provide explanation of the unknown and to mitigate its undesirable consequences. Rather than seek particular sources of particular evils, rather than acknowledge their own limitations and ignorance, these experts assigned the generic name of "witchcraft" to the phenomenology of the unknown. Having a name, they proceeded to found a new professional interest dedicated to its investigation and control.

As the true magnitude of the witch problem became more apparent, the Church enlisted the Inquisition, an applied institution specifically designed to address pressing social concerns. The Inquisition became the growth industry of the day, offering exciting work, rapid advancement, and wide recognition to its professional and technical workers. Its creative and energetic efforts to create a witch-free world unearthed dangers in the most unlikely places; the rates of witch identification, assessment and evaluation soared. By the dawn of the Enlightenment, witches had been virtually eliminated from Europe and North America. Crop failures, disease, and general misfortune had not. And more than half a million people had been burned at the stake, largely 'for crimes they committed in someone else's dreams'. (People are deluded in groups and come to their senses as individuals.)

28-May-09

Stan Young, www.NISS.org

1

We start this lecture with a funny, but somber passage from an essay on risk. After the lecture, we will read the passage again.

Everything is Dangerous, A Controversy

S. Stanley Young

National Institute of Statistical Sciences

May 19, 2009

28-May-09

Stan Young, www.NISS.org

2

We examine statistical analysis strategies of epidemiologists and statisticians using an evaluation method taken from Thomas Kuhn. Kuhn says that it is relatively easy to understand the paradigm of a science by examining their papers, texts and journals. The epidemiology paradigm is to ignore multiple testing. The statistics paradigm is to protect against false discovery.

Abstract

We present evidence of a false discovery rate over 90%. The epidemiology paradigm is “no correction for multiple testing and no sharing of data sets” i.e. “Don’t check on me.”

Some multiple testing mistakes are due to ignorance, but others are intentional, following a (faulty) scientific paradigm; over \$1B of grant/tax money flows to institutions with reproducibility problems revolving around a multiple testing.

28-May-09

Stan Young, www.NISS.org

3

The basic thesis is quite simple. Epidemiologists have as their statistical analysis/scientific method paradigm not to correct for any multiple testing. Also, as part of their scientific paradigm they ask multiple, often hundreds to thousands, questions of the same data set. Their position is that it is better to miss nothing real than to control the number of false claims they make. The Statisticians paradigm is to control the probability of making a false claim. We have a clash of paradigms.

Empirical evidence is that 80-90% of the claims made by epidemiologists are false; their claims do not replicate when retested under rigorous conditions. The net effect of ignoring multiple testing is to exploit randomness.

Beginnings

What is the meaning of life?

What is real?

→ What is reproducible?

Fooled by randomness?

28-May-09

Stan Young, www.NISS.org

4

We leave to the philosophers the meaning of life. Psychologists and physicists can ponder what is real. We and scientists focus on what phenomenon are reproducible. If I conduct an experiment and tell you how I did it, you should be able to get roughly similar results if you conduct a similar experiment.

The effects of randomness are subtle. Humans have to be very vigilant and work very hard not to be fooled by randomness. Some other time, it would be interesting to go into how humans use randomness to fool other humans.

Two books are instructive, *Fooled by Randomness* and *The Black Swan*, by Nassim Taleb.

Epidemiology Recent Claims that do not Replicate

“The reliability of results from observational studies has been called into question many times in the recent past, with several analyses showing that well over half of the reported findings are subsequently refuted.” JNCI, 2007

1. Calcium + VitD for bone breaking
2. Hormone replacement therapy for dementia, CHD, breast cancer, stroke
3. Vitamin E for CHD
4. Fluoride for vertebral fractures
5. Diuretic in diabetes patients for mortality
6. Low fat diet for colorectal cancer and CHD, breast cancer
7. Beta Carotene for CHD
8. Growth hormone for mortality
9. Low dose aspirin for stroke, MI, and death
10. Knee surgery and pain
11. Statins for cancer and mortality
12. Wound dressing on healing speed

Current
Count
1/ 30, 3% !!

28-May-09

Stan Young, www.NISS.org

5

We give the “punch line” first. Claims coming from observational studies fail to replicate in randomized clinical trials 97% of the time! The situation is shocking. We have a scandal.

The NIH has funded a large number of randomized clinical trials testing the claims coming from observational studies. Of 20 claims coming from observational studies only one replicated when tested in RCT. The overall picture is one of crisis.

My current count (very informal data collection) is that only 1 claim of 30 tested is confirmed. This is a 97% failure rate.

Vitamins E and C in the Prevention of Cardiovascular Disease in Men

The Physicians' Health Study II Randomized Controlled Trial

<u>Event</u>	<u>Vit E</u>	<u>Vit C</u>
Major CV Event	NS	NS
Total M. Infarction	NS	NS
Total Stroke	NS	NS
CV Mortality	NS	NS

“no support for the use of these supplements for the prevention of cardiovascular disease in middle-aged and older men.”

JAMA 2008 300, 2123ff

28-May-09

Stan Young, www.NISS.org

0/ 8

6

Vitamins E and C have been repeatedly given rise to claims in observational studies. In a RCT Vits E and C are 0 for 8 on replicating claims.

Statistical Fun and Games

PROCEEDINGS
OF
THE ROYAL
SOCIETY **B**



Proc. R. Soc. B
doi:10.1098/rspb.2008.0105
Published online

You are what your mother eats: evidence for maternal preconception diet influencing foetal sex in humans

Fiona Mathews^{1,*}, Paul J. Johnson² and Andrew Neil³

Claim : Females eating breakfast cereal are
more likely to have a male children !!

28-May-09

Stan Young, www.NISS.org

7

The claim is that women who eat cereal are more likely to have a boy child.

Stop and think a little bit.

A moments thought says, "How can what the woman eats have any influence over the sex of the child, as gender is determined by the X or Y gene coming from the father?!"

What is going on here?

Study Design

1. Three time periods around conception.
2. 133 food item survey.
3. Various composite variables.
4. 740 births.

28-May-09

Stan Young, www.NISS.org

8

Massive number of questions are at issue. $3 \times 133 = 399$ questions at issue. The authors ignore period 3. A few questions had no answers. The net effect is 262 questions are at issue. (The authors also examined some composite questions. It is not all that easy, even when the authors are writing clearly to count the number of questions at issue. For example, should we really ignore period three. Had the authors seen something dramatic in period three, they might have come up with an explanation.)

Young, Bang, Oktay

Comment

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

Implausible biology.

Many questions under consideration.

Complex statistical analysis strategy.

Most claims from observational studies fail on re-test.

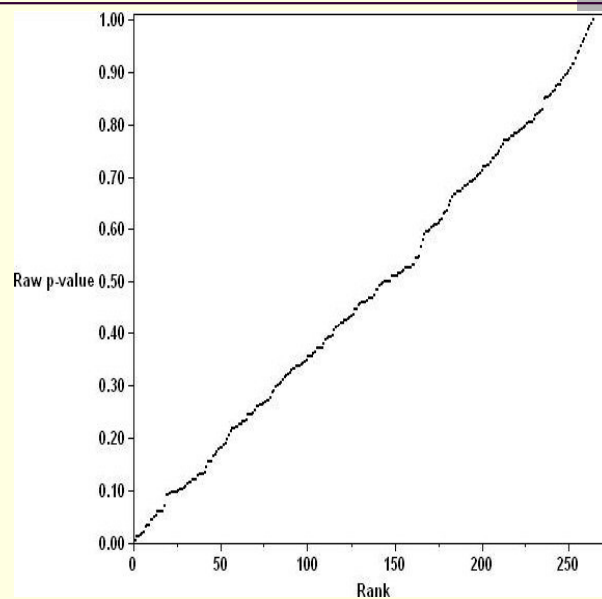
28-May-09

Stan Young, www.NISS.org

9

Heejung Bang noticed in the popular press the claim that eating cereal increased the odds of a boy child. She thought the claim crazy. She also noted that in Asian societies having a male child is a big deal, i.e. there could well be consequences. She asked me to look at the paper. She asked a human fertility expert to join the team. We decided to ask the authors for the data set. Getting observational study data sets can be difficult. In this case the authors had signed an agreement with the journal to make the data available. After some negotiation between the journal and the authors, we were given the data set along with the SAS code the authors used in their analysis. The journal gets an A+ for helping get to the bottom of the authors' claims.

P-value plot of 262 questions



28-May-09

10

In our first analysis, we computed 262 t-tests and plotted the resulting ordered p-values versus the integers giving a p-value plot, Schweder and Spjøtvoll (1982); see Figure 1. Some explanation: Suppose we statistically test ten 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 to 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 data set we have 262 p-values and the plot of the ordered p-values against the integers, 1, 2, ... 262 is essentially linear. This analysis indicates that the data is completely random. The small p-values in the lower left of the figure can be attributed to chance.

Conclusion

Randomness
+
Multiple questions
+
Your tax dollars
=
Fooling by randomness

28-May-09

Stan Young, www.NISS.org

11

Essentially, the man on the street and the popular press have only the most rudimentary understanding of randomness.

Multiple testing is difficult to get your arms around, even for bright, educated people. How does one seemingly unrelated question have anything to do with another. The key is that ALL the Questions are initially at issue. After the analysis is done on all the questions, the p-values are ordered and we look at the smallest ones. We are looking at order statistics and we get fooled by thinking that we are looking at a single, pre-planned comparison.

Many scientists get grants via papers with p-values <0.05. They produce useless papers, duping the public and perhaps themselves as well, while needlessly alarming everyone over findings that will not replicate.

Again, Nassim Taleb is to be congratulated on writing two popular books on randomness. Maybe his next book will be "Exploitation of Randomness!"

Fooled by Randomness Fooling with Randomness

S. Stanley Young

National Institute of Statistical Sciences

Young@niss.org, 919 685 9328

28-May-09

Stan Young, www.NISS.org

12

We now start the formal talk.

It seems rather strange that things bounce around, randomness, and that rationalizations get hung on essentially random events. Nassim Taleb, with his largely personal quest to understand randomness, has done a great service to society by highlighting randomness by writing two books, *Fooled by Randomness* and *The Black Swan*. The world is not so rational as we thought.

I worry that we manufacture “Black Swans that Aren’t”. A random thing happens and humans craving an explanation, pin a story on randomness. Others go one better, using randomness to support their point of view. Doing so, we create so much needless alarm.

Proof : Every study is positive

1. Bias

2. Multiple testing

3. Multiple model searching

Any or all will lead to essentially all observational studies being positive!

28-May-09

Stan Young, www.NISS.org

13

Unless the statistical analysis of observational studies is carefully done, every study will have one or more positive effects.

The word bias covers a lot of sins. Unmeasured confounders. Measure, but unused confounders. Modeling bias – run hundreds of models and select the one you like.

Multiple testing is really quite simple. Ask a lot of questions and only report the ones you want to. Authors can be very clever in hiding multiple testing.

With large complex data sets, there are a number of options available during analysis. These options can be explored until a combination is found that gives a p-value < 0.05 . With complex data sets this is relatively easy. Authors will try this and that until they get a p-value < 0.05 . Some naively believe that $p < 0.05$ means real or they rely on that belief among enough readers to get their paper published.

Editors and referees need to be vigilant to multiple testing. It is a readers beware world.

There is the political problem that to be published a paper must have a claim.

First, Bias

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \dots + \beta_p X_{pt} + \varepsilon$$

$$Y_c = \beta_0 + \beta_1 X_{1c} + \beta_2 X_{2c} + \beta_3 X_{3c} + \beta_4 X_{4c} + \dots + \beta_p X_{pc} + \varepsilon$$

$$\Delta_{t-c} = (\bar{Y}_t - \bar{Y}_c) = \beta_1 (\bar{X}_{1t} - \bar{X}_{1c}) + \beta_2 (\bar{X}_{2t} - \bar{X}_{2c}) + \dots + \beta_p (\bar{X}_{pt} - \bar{X}_{pc}) + (\bar{\varepsilon}_t - \bar{\varepsilon}_c)$$

$$\Delta_{t-c} - [\text{known confounders}] = \beta_1 + [\text{unknown confounders}]$$

28-May-09

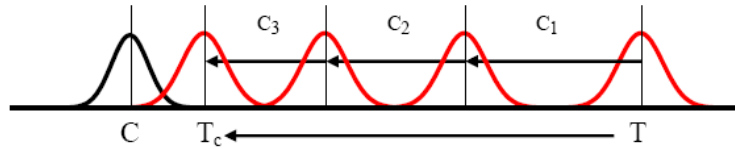
Stan Young, www.NISS.org

14

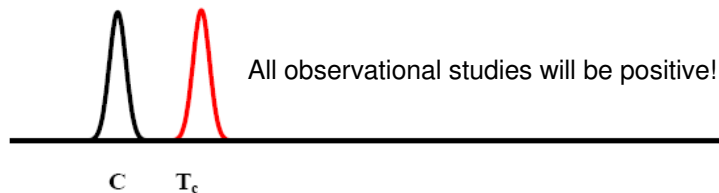
Consider a linear model for a treated individual and a control individual. Let X_{1t} indicate treatment and take the value 1 and X_{1c} indicate no treatment. The remaining X 's are covariates. If we average all the treated and control individuals and subtract the two resulting equations, we get a delta for the difference between treated and control individuals. Now if we move all the known confounders to the left of the equation, we take out the effect of the known confounders. Unknown confounders are still confounded with the treatment difference and can confuse the interpretation of the data.

Residual bias: observational studies

(a) Use confounding variables to reduce bias.



(b) As n get large the standard error of the mean gets small.



28-May-09

Stan Young, www.NISS.org

15

In an observational study, most typically there is a difference between the control and treated groups. As confounding variables are removed, the treatment effect moves toward the control group. If there are unknown or unmeasured confounders the treatment groups remain separated.

Observational studies are getting larger. As sample size gets larger the standard error of the mean gets smaller so that small bias can result in a statistically significant claim, false discovery, that is the result of bias not treatment.

The rule of thumb 5 years ago was that if the risk ratio, RR, was not larger than 2 then any observed effect could be the result of confounders and it was improper to make any claims. A RR has to be larger than 2 to be admissible in federal court.

A small survey was taken of journal editors. Epidemiology journals now have no requirement that a risk ratio be greater than 2 to be taken seriously.

Bias

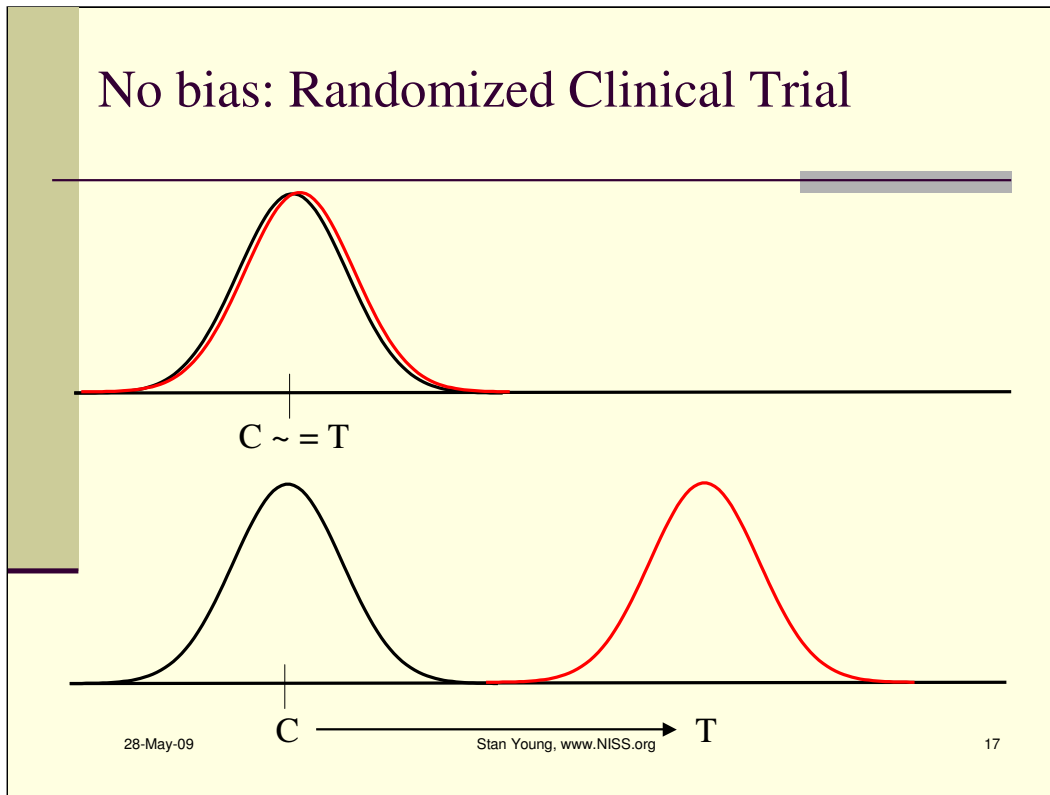
Observational studies are likely to have residual bias.

As the sample size gets large, residual bias will likely lead to “statistical significance”.

Bias is not expected to go to Zero as sample size increases.

In summary, observational studies are likely to have residual bias and if the study is large there is a high likelihood that there will be “statistical significance”.

No bias: Randomized Clinical Trial



For RCT, through randomization the effects of bias are largely, but not completely, removed.

If treatment has an effect it will move the distribution of the treated patients away from the control patients. If the effect is large enough and if the sample size is large enough, the treatment effect will be detected.

10-sided dice experiment

1. Pair up, one rolls, one records the p-value.
2. If you get 0-0, then roll again to get 0.00xx
3. Circle all p-values < 0.05 (04, 03, 02, 01, 00)
4. Note any col or row with multiple significances.
5. Hold your sheet up when finished.

28-May-09

Stan Young, www.NISS.org

18

Mathematical treatment of randomness can be rather intellectual without giving much sense of what is going on at a visceral level. Real experiments help get a feel for randomness. Ten-sided dice have the numbers 0, 1, ..., 8, 9 on ten faces of each die. Using two colored dice, you can get p-values, 0.xy using one die for 0.x_ and the other for 0._y. when rolled, if you get 04, 03, 02, 01, 00, then that can be taken as getting a p-value < 0.05 .

<u>MedCondition</u>	<u>YoungFemale</u>	<u>YoungMale</u>	<u>OldFemale</u>	<u>OldMale</u>
1. Angina				
2. Arthritis				
3. Asthma				
4. Cancer				
5. C. Bronchitis				
6. CHD				
7. Emphysema				
8. Heart Attack				
9. Liver Disease				
10. Stroke				
11. Thyroid D.				
12. Diabetes				
13. H. LDL				
14. L. HDL				
15. C React Protein				

28-May-09

Stan Young, www.NISS.org

19

Our study consists of 60 statistical tests, four age/gender categories and 15 medical conditions. The dice are rolled 60 times and the p-values are entered into the cells of the data table. With 60 questions, there is only about a 5% chance that your study will not have a significant p-value.

How do you determine if a claim from an observational study is plausible?

Questions for Observational Study (all conditions must be met)

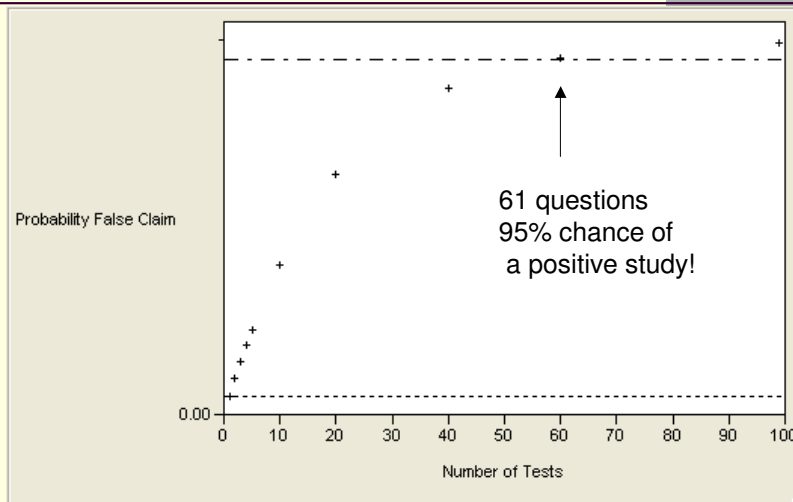
1. There is a viable biological explanation, not post hoc.
2. The data set and analysis code has been examined by an independent group.
3. The data set and analysis code is publicly available to anyone.
4. Any measures of uncertainty of the findings are multiplicity adjusted.
5. The finding has been replicated.

What you can do?

1. Look for crazy claims in press.
2. Get original paper and press release.
3. Count questions and models (not as easy as it sounds).
4. Ask for data (Most likely you will not get it).
5. Write Letter2Editor to journal. Comment on multiple testing or "Trust me" science.

Remember: People are deluded in groups and come to their senses as individuals. Individuals need to recognize problems and make an effort to correct them.

How do you get a “ $p < 0.05$ ”? Ask lots of questions.



The authors start with $3 \times 133 = 399$ potential questions.

28-May-09

Stan Young, www.NISS.org

20

Just ask a lot of questions in a study and you are very likely to get a statistically significant result by chance alone.

If 61 independent questions are asked in an experiment there is a 95% probability of at least one “statistically significant” result.

A rule of thumb is to multiply any reported p-value by the number of questions under consideration. To be statistically significant after this adjustment, the resulting adjusted p-value should be below 0.05.

So in a large, complex study, just ask a lot of questions.

A multiple testing/modeling train wreck

Association of Urinary Bisphenol A Concentration With Medical Disorders and Laboratory Abnormalities in Adults *JAMA. 2008;300(11):1303-1310*

1. 275 chemicals
2. 32 medical outcomes
3. 10 demographic covariates

$$275 \times 32 = 8800 \times 2^{10} = \sim 9 \text{ million}$$

This train wreck is in progress!

28-May-09

Stan Young, www.NISS.org

21

We are in a target rich environment for false claims. Note that bisphenol A is a critical industrial chemical. It would be a tragedy for this chemical to be restricted/removed/replaced over a false positive claim.

To the Editor: From a cross-sectional analysis of urinary chemical concentrations and health status in the general US adult population, Dr Lang and colleagues reported that BPA was associated with cardiovascular diagnoses, diabetes, and abnormal liver enzyme concentrations. However, the potential for false positives, briefly mentioned but not analyzed, is substantial when the complete Centers for Disease Control and Prevention (CDC) design is examined.

The CDC NHANES (2003-2004) measured 275 environmental chemicals and a wide range of health outcomes. Although the study by Lang et al focused on 1 chemical and 16 health outcomes (8 patient-reported medical outcomes and 8 clinical chemistry measurements), counting to determine how many questions were at issue and in how many ways these questions can be statistically analyzed is important.

Focusing only on the health outcomes selected by the authors, the analysis forms a 16275 composite set of questions. However, there are more than 8 ways that the medical outcomes can be examined since 2 of the outcomes have subgroups, any 1 or combination of which could result in an association. Likewise, there are more than 8 ways the clinical measurements can be examined because additional measurements and derived outcomes were reported. Overall, we counted 32 possible outcomes.

From the perspective of the complete CDC study design, there are $32 \times 275 = 8800$ questions at issue. In addition, there is a large list of possible confounder variables; we counted 10. The authors used 2 regression models to adjust for confounders, but with 10 confounders, there are 1024 possible different adjustment models. Considering the complete list of questions at issue and confounders, the model space could be as large as approximately 9 million models.

Given the number of questions at issue and possible modeling variations in the CDC design, the findings reported by the authors could well be the result of chance. The authors acknowledged as much for only 16 questions for BPA alone, and we amplify their warning by pointing out the conceptually much larger CDC grand design. There could easily be a flood of articles reporting chance results. We note that *JAMA* recently published an article reporting an association between arsenic and diabetes using the same database.

We think it is a good time to step back and consider the entire CDC study for the large, planned study that it is and develop a statistical analysis strategy that takes into account the large number of questions at issue.

End of proof

Combination of residual bias,
multiple testing
multiple analysis
of complex study

You are a winner – every study is positive!

Indiscriminant multiple testing and/or residual bias (and large data sets) can lead to essentially every study having one or more significant effects.

Epidemiology Science Paradigm

(Thomas Kuhn – historical, what they do.)

1. Examine many questions in non-randomized studies.
2. No adjustment for multiple testing. (Appear to follow Karl Popper, asserting that these are pre-specified, falsifiable hypotheses.)
3. Within each question, use alternative analysis strategies.
4. From the many claims, select one or a few for reporting. (Use subject matter knowledge to make a final list of claims.)
5. Impose no standard on the magnitude of an effect deemed reportable
 - a. The unadjusted p-value is <0.05 .
 - b. A plausible explanation of the effect can be proposed.
6. Although the search for possible claims is essentially retrospective, the writing of the claims should be as close as possible to Popper “we tested this pre-planned hypothesis”. See Taleb.

28-May-09

Stan Young, www.NISS.org

23

So, using the historical approach of Kuhn, what is the current operable epidemiology paradigm?

There is the very human characteristic to find a rational explanation for observations. See in particular, Fooled by Randomness and the Black Swan by Taleb.

To an outsider looking in: Paper Writing Paradigm*

1. There will be no mention of multiple testing.
2. There will be no enumeration of the number of questions under consideration.
3. There will be no pre-experiment definition of the statistical analysis strategy, i.e. no statistical protocol.
4. There will be no public posting or sharing of data sets.
5. There will be no criticism of the statistical methods of others with respect to multiple testing.

Adjustment for multiple testing is not part of the paradigm.

28-May-09

Stan Young, www.NISS.org

24

* A very few counter-examples exist .

Over 90% of Epidemiology papers follow this paper writing paradigm, so following Kuhn, we conclude that correction for multiple testing is not part of the scientific paradigm of epidemiologists.

At best, claims coming from most epidemiology studies should be considered “hypothesis generating.”

If no data set is publicly available, it is “trust me” science.

Two Paradigms

1. Every statistics student learns about Type 1 error and multiple testing.
2. Epidemiology students are taught

*No adjustments are needed for multiple comparisons.
Rothman: Epidemiology 1990, 1:43–46.*

Epidemiologists paradigm : test everything and sustain any level of type 1 errors not to miss anything.

See also, Vandembroucke, PLoS Med (2008), Young, IJE (2009).

28-May-09

Stan Young, www.NISS.org

25

Here are the two paradigms under discussion. Statistics is aimed at how to efficiently obtain knowledge of the world. There is randomness in the world so that needs to be taken into account in the knowledge gathering process. Every statistician or person that takes a statistics course taught by a statistics department understands the risk of making a false claim based on a statistical analysis.

Epidemiologists understand Type 1 error and false positives. Their operable scientific paradigm is not to control for false positives.

Vandembroucke restates and agrees with Rothman. Many leading epidemiologists “sign on” to his paper.

Leaving no trace

Usually these attempts through which the experimenter passed, don't leave any traces; the public will only know the result that has been found worth pointing out; and as a consequence, someone unfamiliar with the attempts which have led to this result completely lacks a clear rule for deciding whether the result can or can not be attributed to chance.

28-May-09

Stan Young, www.NISS.org

Shaffer, 2007 26

Quite important. The epidemiologists, in effect, assume that every question is independent. It is within their paradigm to ask many questions of a data set and report positive findings in separate papers. Most often they do not say how many questions were under consideration and they often do not give details of their statistical analysis.

Doublethink, George Orwell

The FDA insists on replication for any efficacy claim.

The FDA ignores multiple testing for side effects.

The FDA is trying to address observational studies, OMOP
but

OMOP science advisory board: 3 MPH
1 Statisticians.

OMOP is funded by pharma, \$22M

28-May-09

Stan Young, www.NISS.org

27

Doublethink is the ability to hold two contradictory thoughts in your head and believe them both. The FDA insists on multiple testing correction and replication for efficacy claims, but requires neither for side effects.

As more medical claims are being made based on observational studies, the FDA is trying to address these claims. Hopefully the effort will suitably address bias, multiple testing, and multiple model building.

*References

Pocock SJ, Collier TJ, Dandreo KJ, et al. Issues in the reporting of epidemiological studies: a survey of recent practice. *BMJ*. 2004;329:883-888.

Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. 2005;294:218-228.

Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology*. 1990;1:43-46.

Shapiro, S. Looking to the 21st century: have we learned from our mistakes, or are we doomed to compound them? *Pharmacoepidemiology and Drug Safety* 2004;13:257-265.

Austin PC, Mamdani MM, Juurlink DN, Hux JE. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *J Clin Epidemiol* 2006;59:964 - 969.

28-May-09

Stan Young, www.NISS.org

28

Pocock said that epidemiology is in crisis in 2004. The epidemiologists appear to be in denial.

Ioannidis, *JMAM* 2005, points out the ~80% false discovery rate of epidemiology.

Rothman says no correction for multiple testing is necessary. This is the current epidemiology paradigm.

Shapiro will have nothing of it and points to an example of a false positive result of 30 years ago from which the epidemiologists seemed to learn nothing.

Austin uses a humorous example to show how false positives can result from multiple testing. This is a great paper, a must read.

*References (2)

Sesso, HD, Buring, JE, Christen, WG, et al. Vitamins E and C in the prevention of cardiovascular disease in men. JAMA. 2008;300, 2123-2133.

Peto, R, Emberson, J, Landray, et al. Analysis of cancer data from three Ezetimibe trials. NEJM. 2008;358, 1357-1366.

Vandenbroeke, JP. Observational research, randomised trials, and two views of medical science. PLoS Medicine 2008;5, e67, 0339-0343.

Young, SS (2008) Everything is dangerous: a controversy.
http://www.niss.org/talks/Young_Safety_June_2008.pdf

The Peto paper is important for its analysis strategy. You have to actually commit resources to replication.

Credits

Heejung Bang, Cornell University

Kutluk Oktay, New York Medical College

Ya-Lin Chiu, Cornell University

Stuart Taylor, Royal Society

Heejung Bang is a statistician. Kutluk Oktay is practicing physician and an expert in human fertility. Ya-Lin Chiu did the statistical analysis of the Mathews data, p-value plot and multiplicity adjusted p-values. Stuart Taylor is one of the many supporting editors at the journal of the Royal Society. These editors worked hard to understand our statistical objections to the Mathews paper and get us the data set. Again, they get an A+ for their efforts. They were great to work with.

Post processing

From WITCHES, FLOODS, AND WONDER DRUGS: HISTORICAL PERSPECTIVES ON RISK MANAGEMENT, by William C. Clark. "For several centuries spanning the Renaissance and Reformation, societal risk assessment meant witch hunting. Contemporary accounts record wheat inexplicably rotting in the fields, sheep dying of unknown causes, vineyards smitten with unseasonable frost, human disease and impotence on the rise. In other words, a litany of life's sorrows not very different from those which concern us today.

The institutionalized expertise of that earlier time resided with the Church. Then, as now, the experts were called upon to provide explanation of the unknown and to mitigate its undesirable consequences. Rather than seek particular sources of particular evils, rather than acknowledge their own limitations and ignorance, these experts assigned the generic name of "witchcraft" to the phenomenology of the unknown. Having a name, they proceeded to found a new professional interest dedicated to its investigation and control.

As the true magnitude of the witch problem became more apparent, the Church enlisted the Inquisition, an applied institution specifically designed to address pressing social concerns. The Inquisition became the growth industry of the day, offering exciting work, rapid advancement, and wide recognition to its professional and technical workers. Its creative and energetic efforts to create a witch-free world unearthed dangers in the most unlikely places; the rates of witch identification, assessment and evaluation soared. By the dawn of the Enlightenment, witches had been virtually eliminated from Europe and North America. Crop failures, disease, and general misfortune had not. And more than half a million people had been burned at the stake, largely 'for crimes they committed in someone else's dreams'. (People are deluded in groups and come to their senses as individuals.)

28-May-09

Stan Young, www.NISS.org

31

One can make the case that statistical analysis of observational studies are the "witch hunts" of our time.