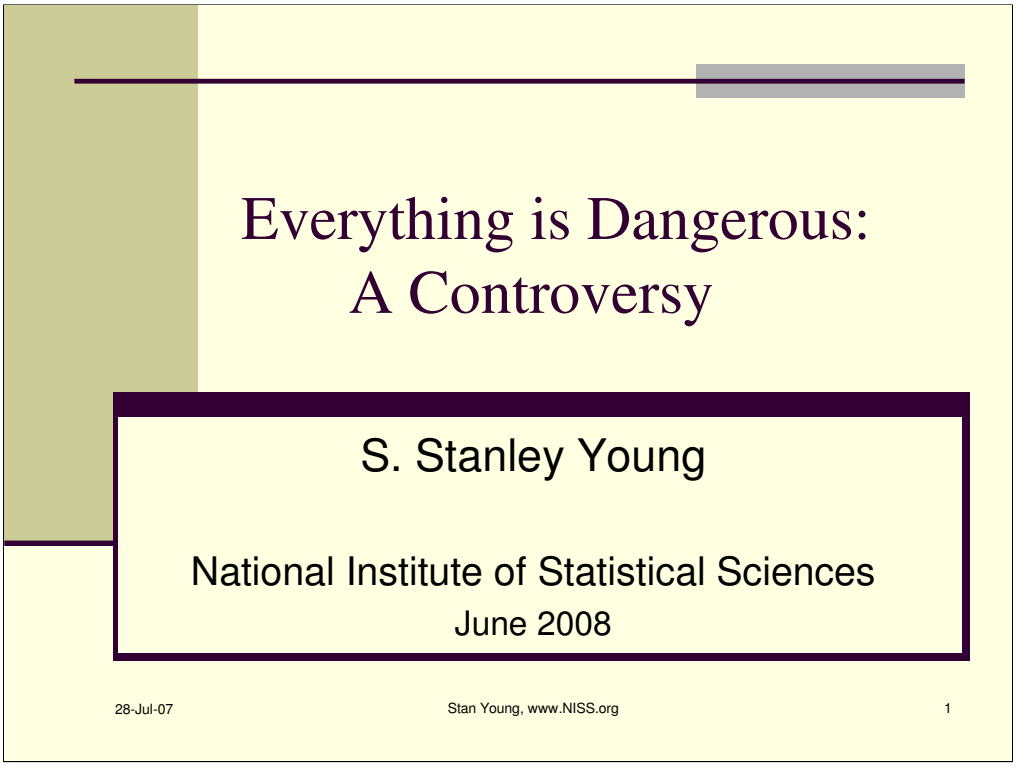


We examine statistical analysis strategies of epidemiologists and statisticians using an evaluation method taken from Thomas Kuhn. Kuhn says that it is relatively easy to understand the paradigm of a science by examining their papers, texts and journals. The epidemiology paradigm is to make no correction for multiple testing. The statistics paradigm is to protect against chance false discovery.

Let me say at the beginning, I think medical observational studies are important and can be analyzed in a matter that claims are dependable. Epidemiologists are well-versed in statistics and are capable of defending their paradigm.

Abstract

Some multiple testing mistakes are due to ignorance (how often are you asked to re-examine the data to see if something can be found?), 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. Statisticians need to understand other scientists' paradigms. It serves neither society nor our profession to ignore multiple testing controversies. At a minimum we need to protect the integrity of our profession. We present evidence of a false discovery rate over 80%. We present survey of journal editors on multiple testing that support the epidemiology paradigm of no correction for multiple testing and not sharing of data sets.

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, of 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; these claims do not replicate when retested under rigorous conditions.

Epidemiology

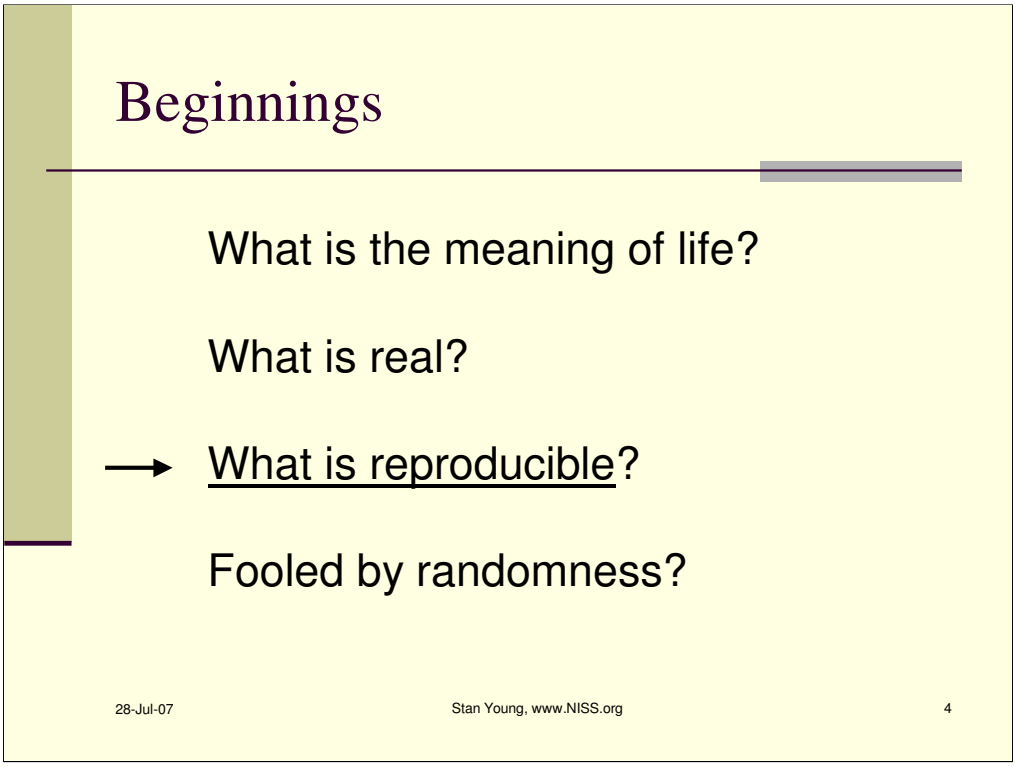
Recent Claims that do not Replicate

"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

1/ 20, 5% !!

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.

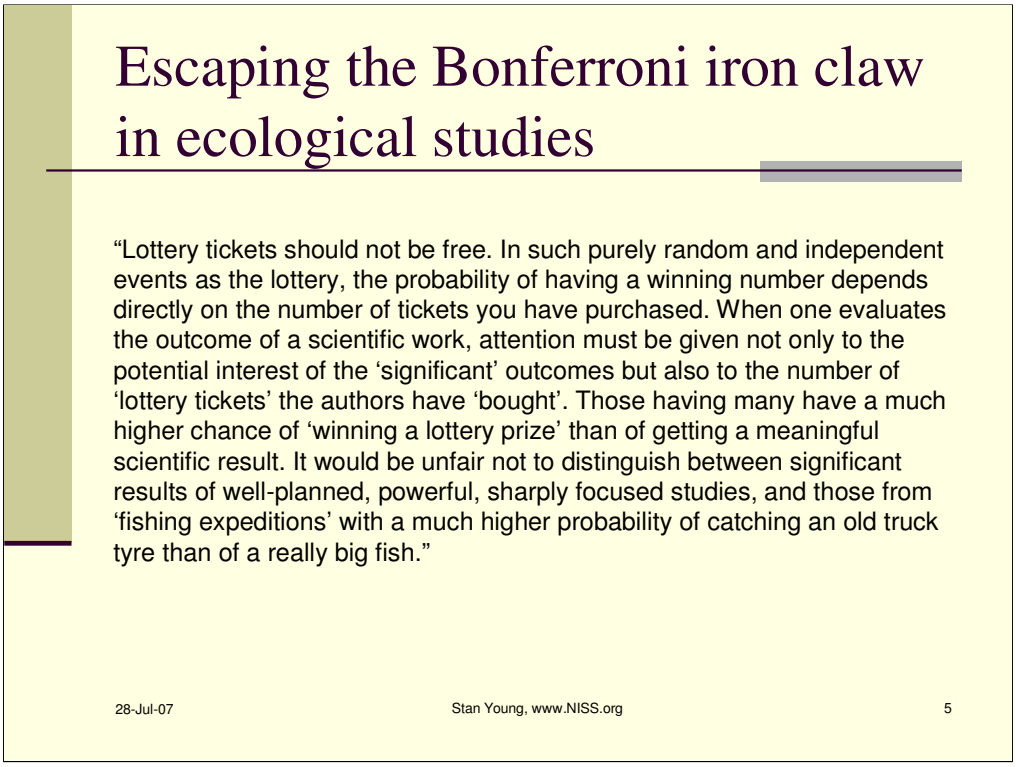


We leave to the philosophers the meaning of life. Psychologists and physicists can ponder what is real. We and scientists focus on what phenomenon are reproducible. If I conduct an experiment and tell you how I did it, you should be able to get roughly similar results if you conduct a similar experiment.

The effects of randomness are subtle. Humans have to be very vigilant and work very hard not to be fooled by randomness.

See two books by Nassim Taleb, Fooled by Randomness and The Black Swan.

Some other time, it would be interesting to go into how humans use randomness to fool other humans.



Multiple testing is not just a problem of epidemiology. I use epidemiology as an example as they are not correcting for multiple testing as part of their scientific paradigm. They understand multiple testing. They are not doing what they are doing through ignorance. See for example, Vandenbroucke, PLoS Med (2008).

Clinical trials and Genetics/Genomics are two sciences that take multiple testing seriously.

Ioannidis in Journal of the American Medical Association examined highly cited medical trials, non-randomized and randomized, and found that claims coming from non-randomized trials failed to replicate or the claimed effect was dramatically smaller when the claim was tested a second time. Ioannidis noted that claims coming from randomized medical trials failed to replicate about 20% of the time.

Stuart Pocock in BMJ catalogues the current problems with the reporting of epidemiology studies. There are so many problems that it is difficult to say that multiple testing is the largest problem. I think Pocock underestimates the multiple testing problem when he says the false discovery rate of epidmiology is on the order of 20%. On the other hand, it is relatively easy to fix the multiple testing problems: Copy the statistical strategies used in randomized clinical trials.

Outline

1. The question.
2. Two proofs
3. Two paradigms.
4. Crisis?

The motivation of this lecture is that effects found in non-randomized (epidemiology) medical studies are failing to replicate when tested in randomized clinical trials. The false discovery rate of epidemiolgy studies might be considered excessive, ~80-90%. We present two paradigms for the analysis of non-randomized studies. The epidemiology paradigm is to test many questions and with no adjust for multiple testing. The statistics paradigm is to correct the analysis for the number of questions asked controlling the false positive rate at a fixed level, usually 5%. Is there a crisis? When an important science, epidemiology, has a false discovery rate of 80-90%, there appears to be a crisis.

Statistical Fun and Games

PROCEEDINGS OF THE ROYAL SOCIETY B

FirstCite® e-publishing

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. Johnson[2] and Andrew Neil[3]

Example: 54 p-values, smallest is 0.003.

54 x 0.003 = 0.162.

Over 51,000 Google hits!

The claim coming from this paper is not significant when multiple testing is taken into account. The paper was wildly popular with the public press. It was even written up in the Economist. After much discussion, intervention by the editor, and the signing of rather restrictive legal document, it appears that we will get the data set from the authors.

Proof : Every study is positive

1.Bias

2.Multiple testing

Either or both lead to all observational studies being positive!

Unless the statistical analysis of observational studies is carefully and conscientiously conducted, every study will have one or more statistically significant effects. We chose to focus on two statistical issues, bias and multiple testing.

First, Bias

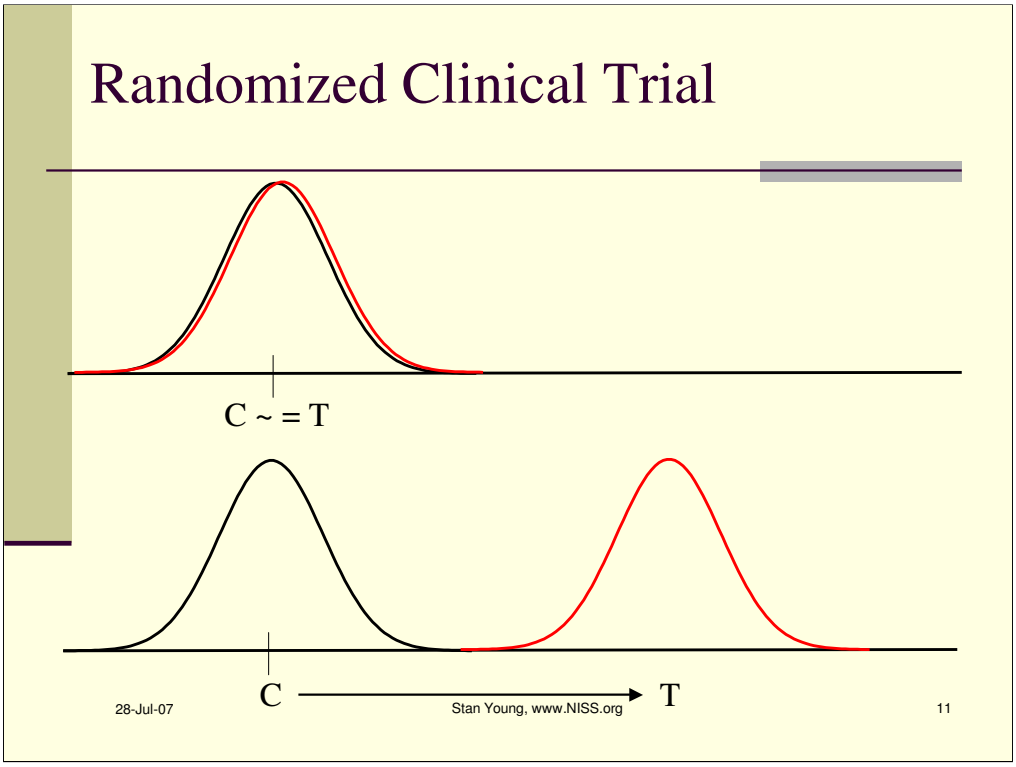
$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \ldots + \beta_p X_{pt} + \varepsilon$$

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

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

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

Consider a linear model for a treated individual and a control individual. Let X1t indicate treatment and take the value 1 and X1c 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.

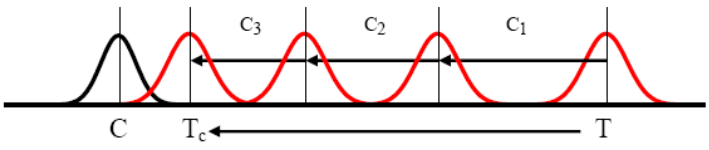


For RCT, through randomization the effects of bias are largely, but not completely, removed. If there is no treatment effect the two distributions are on top of one another.

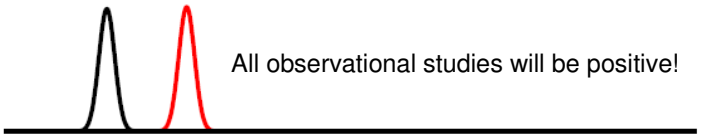
If treatment has an effect it will move the distribution of the treated patients, red, away from the control patients. If the effect is large enough and if the sample size is large enough, the treatment effect will be detected.

Bias reduction in observational studies

(a) Use confounding variables to reduce bias.



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

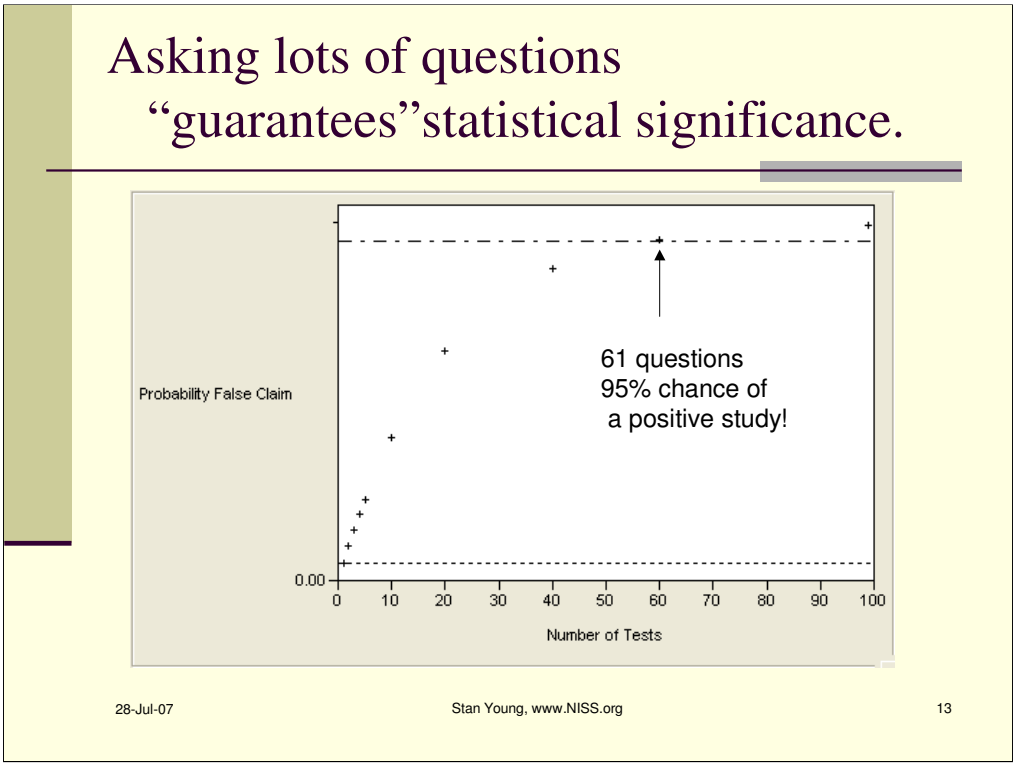


All observational studies will be positive!

In an observational study, most typically there is a difference between the control and treated groups. The groups will differ in important confounding variables, age, income, etc. As confounding variables are mathematically removed, the treatment effect moves toward the control group. If there are unknown or unmeasured confounders, then 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 considered improper to make any claims. A RR has to be larger than 2 to be admissible in federal court. As a point of reference, the RR of smoking is on the order of 8-10.



If you do one statistical test, when nothing is really going on, you will get statistical significance 5% of the time. The methods are designed to have a false positive rate of 5%. Now if you do two independent tests there is just less that a 10% chance that you will have one or more statistically significant results, again by chance alone. The 5% usual false positive rate was rather arbitrarily proposed by R.A. Fisher many years ago and it has become the norm for evaluation of experiments. The trade off is roughly, making a mistake 5% of the time is ok, relative to the cost of experimentation if you require stronger evidence.

The left axis gives the chance of one or more statistically significant results, again assuming that nothing is really going on. The x axis gives the number of independent statistical tests. If 61 independent questions are asked in an experiment there is a 95% probability of at least one "statistically significant" result.

A rough rule of thumb is to multiply any reported "raw" 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.

End of proof

Combination of residual bias,
large sample size and
multiple testing

You are a winner – every study is positive!

3 time points x
133 food items =
399 statistical tests!

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

These authors, among other tests, tested 133 food items at three time periods to give a total of 399 statistical tests of significance. The raw p-value they used to make their claim was 0.029. Any adjustment renders this test not significant.

The interpretation of the adjusted p-value is “the probability that you will see a raw p-value this small given the number of questions under consideration.” If you do a lot of tests you should expect to see some small raw p-values.

Randomized Clinical Trials
Side effects and multiple testing

FDA demands no multiple testing correction for side effects.

What to do?

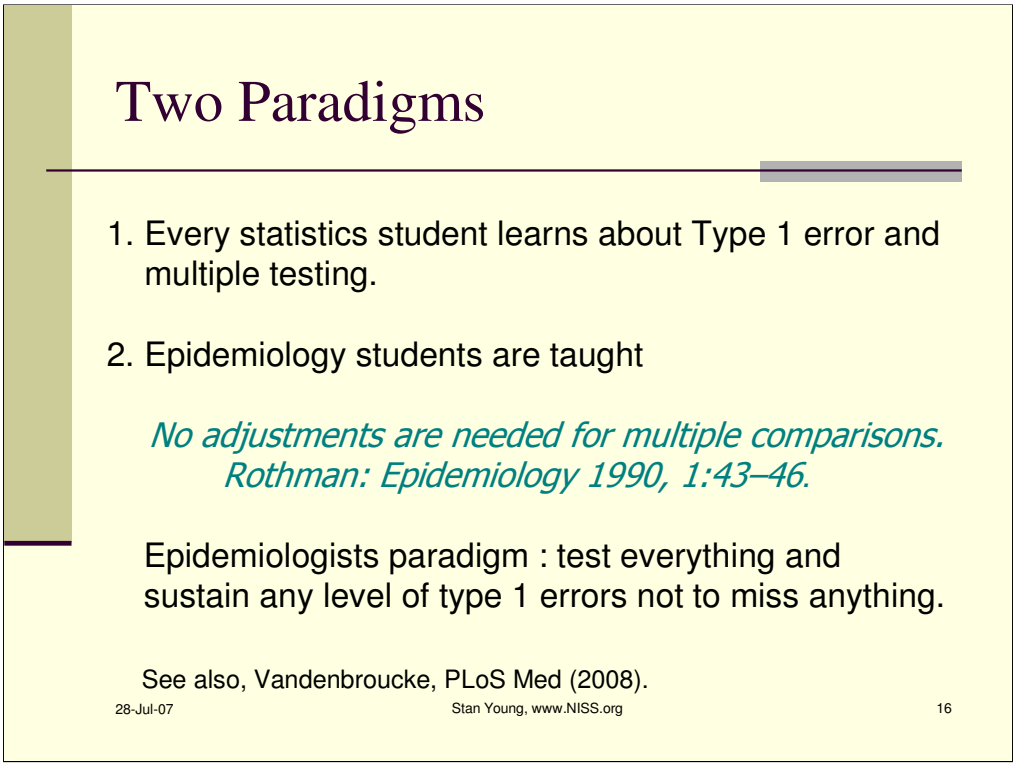
1. Pre-plan/specify categories.
2. Stage analysis.
    a. Analyze trials separately.
    b. Give both raw and adjusted p-values.
    c. p-value plot on combined analysis.
3. Look for beneficial side effects as well.

This is a short excursion to look at randomized clinical trials.

For efficacy there is general agreement that multiple testing has to be controlled.

Side effects are treated very differently (and not consistently).

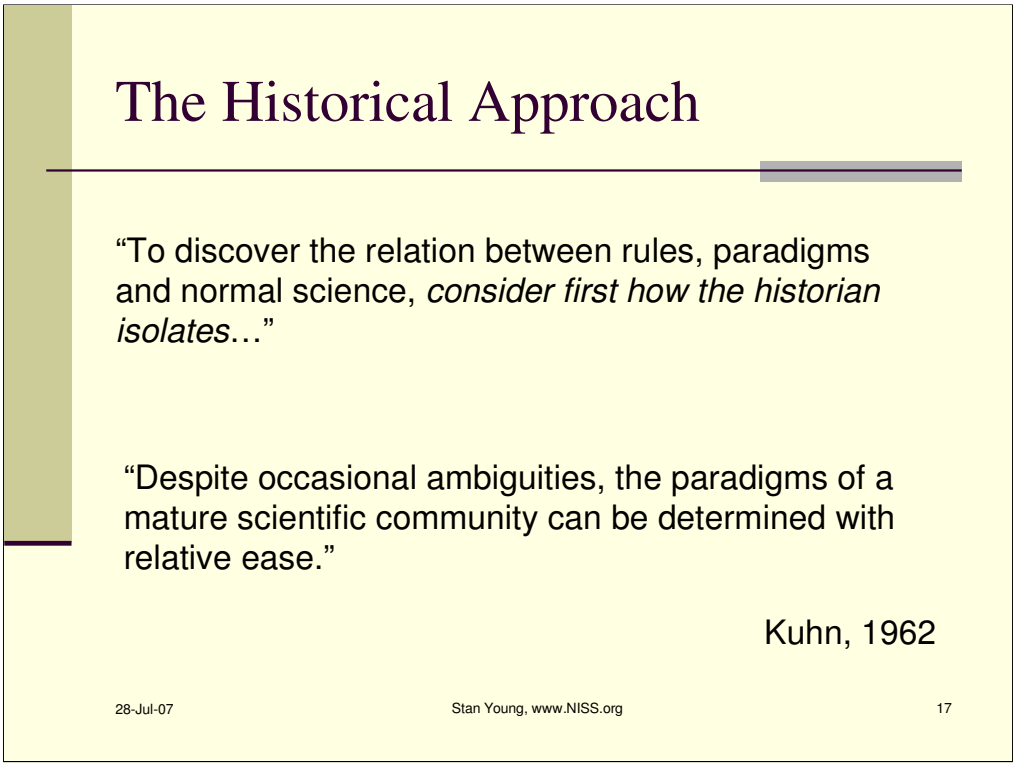
Multiple testing comes up in RCTs with the analysis of side effects. There needs to be a systematic strategy for the analysis of side effects. Make side effect categories. Do not just allow an unspecified list to develop. Phase III trials come in pairs. Analysis the pairs separately and together. Give both unadjusted and adjusted p-values. Look for decreases (benefits) as well as increases in side effects.



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. That probability is controlled at 5%.

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

Vandenbroucke restates and agrees with Rothman. Many leading epidemiologists “signed on” to his paper. Epidemiologists understand statistics and false positives; they chose not to control the false positive rate as part of their scientific paradigm.



In a classic book, Thomas Kuhn describes how paradigms change in science. There is "normal science". Some abnormalities are noticed. There is chaos as these abnormalities are considered in light of existing theory. A new theory/paradigm is developed to explain the abnormalities and then there is a return to normal science, working out the details of the new paradigm.

Kuhn takes a historian's point of view. Examine the text books, the scientific papers and the lectures that teach students the craft. He makes the point that the operable paradigm of a mature science is easy to determine.

Kuhn, T.S. (1962) The Structure of Scientific Revolutions, 3rd Edition, The University of Chicago Press

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.

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

There is the very human characteristic to construct a rational explanation for observations. See in particular, Fooled by Randomness and the Black Swan by Taleb. Unfortunatedly, the smartest people come up with the most plausible rationalizations.

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.

* A 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.

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.

Shaffer, 2007

Quite important. The epidemiologists, in effect, assume that every question is independent, and is to be taken out of the context of the experiment. 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.

See slide 13 again.

Maverick Solitaire

Maverick Solitaire. Given a normal 52-card deck of playing cards, shuffle, and then deal 25 cards. Set aside the rest of the deck. Attempt to arrange the 25 cards into five hands of five cards each, such that each hand is "pat", a flush, a straight, a full house, or four of a kind.

98% on first 100 deals.

Retrospective rationalization. After you get the 25 cards you can arrange them to get 5 perfect hands*. If you are dealt one set of five cards, a perfect hand is very rare.

Once an experimenter examined the data, they can almost always come up with a plausible explanation.

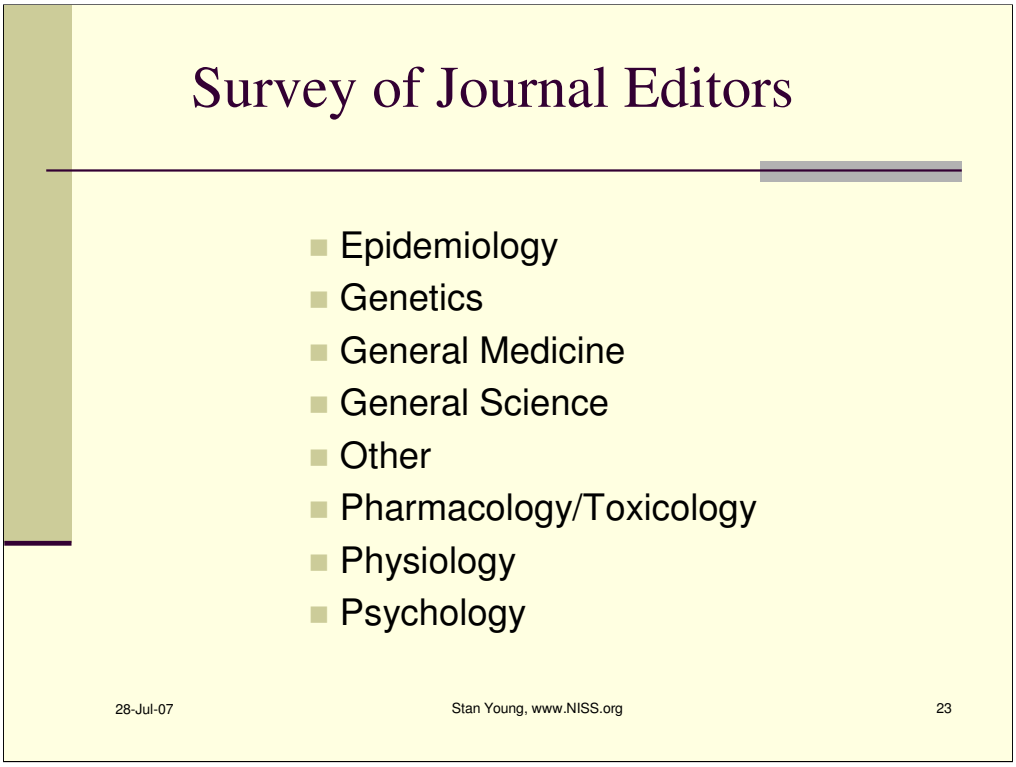
*The term “pat” comes from the gambler patting his cards on the table indicating that he want no additional cards.

Retrospective Rationalization

"In particular, it is important that biologically plausible associations be specified during the design of the study, because it is tempting to construct biologically plausible reasons for observed subgroup effects after having observed them."

Peter Austin

A number of authors make the point that humans are very good at giving a plausible explanation AFTER they see the data. Retrospective rationalization should count for little as it is so easily done.

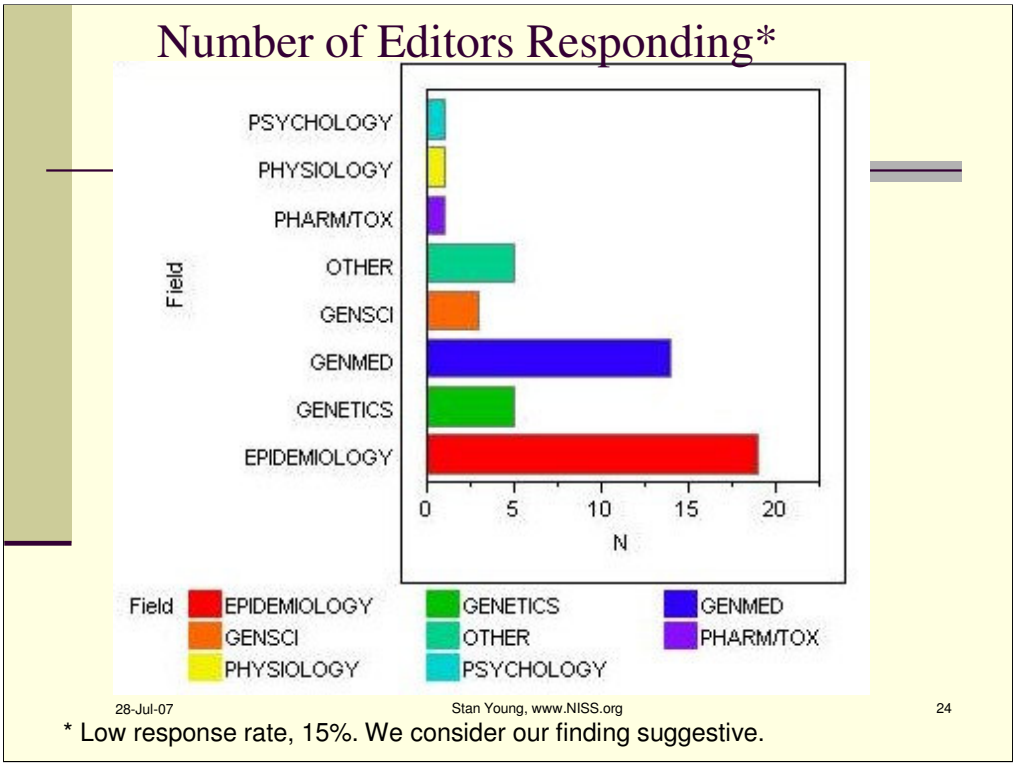


We conducted a survey of editors of science journals asking a number of questions.

We used the “wheat and chaff” technique of putting the key questions in amongst other questions.

The key questions for us were related to the treatment of multiple testing and access to data sets used in publications.

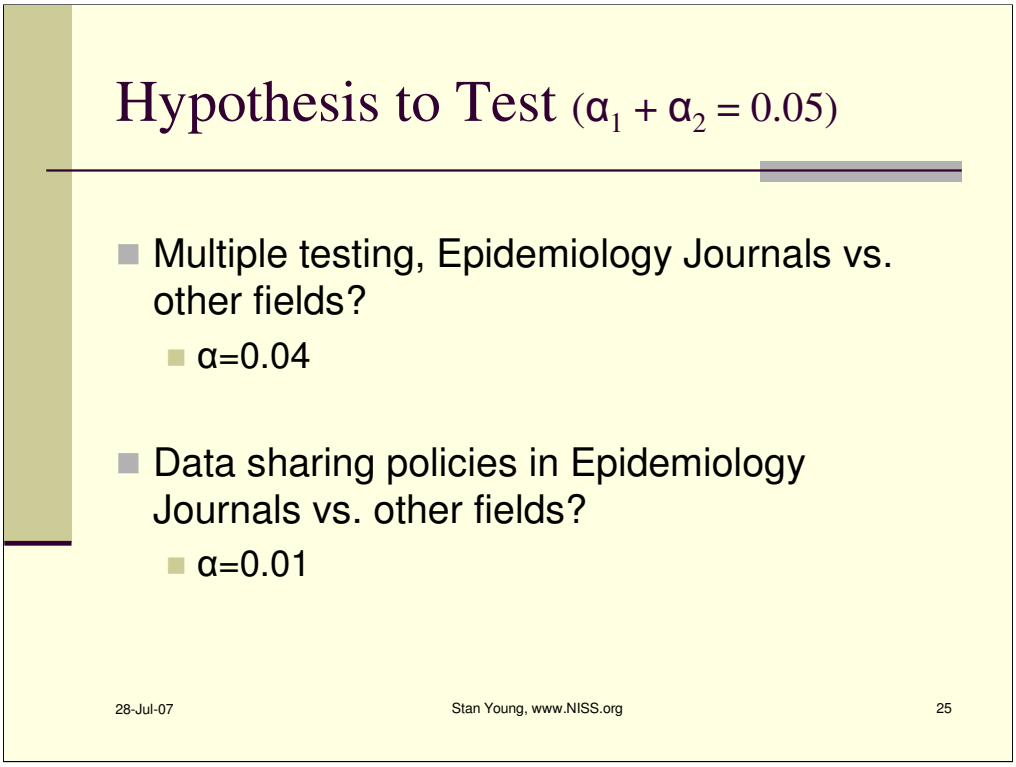
It is a basic tenet of science that scientist should help other scientists evaluate their work. One scientists is suppose to share his data set with another. If epidemiologists do not share their data sets, the reader is left with “trust me” and that is not science. As a point of reference, any RCT that is used for drug approval must be given to the FDA along with the code used to compute the statistical analysis.



We give the number of editors responding.

We intentionally over-sampled epidemiology editors as we wanted to know their opinions on multiple testing and data sharing.

The response rate to the web survey was low so our finding must be regarded as tentative.



We pre-determined how we would do the statistical analysis. We wanted an overall 5% error rate so we allocated 1% to the data sharing question and 4% to the multiple testing question.

We went into the survey expecting that epidemiology editors would require no adjustment for multiple testing and have no policy of making data sets available.

This work was done by Mike Last, as NISS post doc at the time. The questions were carefully constructed and tested on a small number of people in an attempt to be sure the questions and answers were clear.

Multiple Testing, Question 1 Coding

- To what extent should pre-specified study protocols, as opposed to reported results, address multiple testing issues?
  1. No need or requirement for a formal experimental protocol or plan to address multiple testing; authors *may* comment (or not) on multiple testing issues
  2. **Authors *must* comment on any multiple testing issues**
  3. **Authors must specify the number of questions under consideration**
  4. **Author must have a pre-specified statistical testing protocol that adjusts for multiple testing**
  5. Not applicable/rarely arises

Here is the question and possible answers. The contrast is answer 1 vs answers 2,3,4. Note that for randomized trials the scientist must pre-specify the testing protocol.

Multiple Testing Question 2

- How does the journal deal with data sets used for more than one study? For example, some large surveys lead to many papers.
  1. Multiple publications based on analysis of new questions of the same data set is considered appropriate without comment by authors
  2. **Authors should discuss previous uses of the data set**
  3. **Data may be re-used for hypothesis generation, but this is not appropriate for hypothesis testing (p-values or confidence limits)**
  4. Not applicable/rarely arises

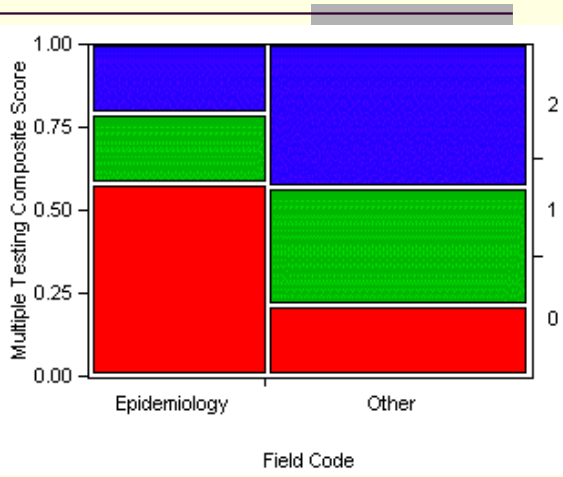
Here is the second question on multiple testing.

Basically, how many times can you use a data set?

A very large epidemiology study might have hundreds to thousands of questions and generate tens to hundreds of papers.

Analysis of Multiple Testing

- Multiple Testing:
  - Ignore double-N/A
  - Score 1 for each bolded answer
  - Epidemiologists:
    Score- 12/19
  - Other:
    Score- 34/28
  - P=0.019, ANOVA
    - General Medicine only good group



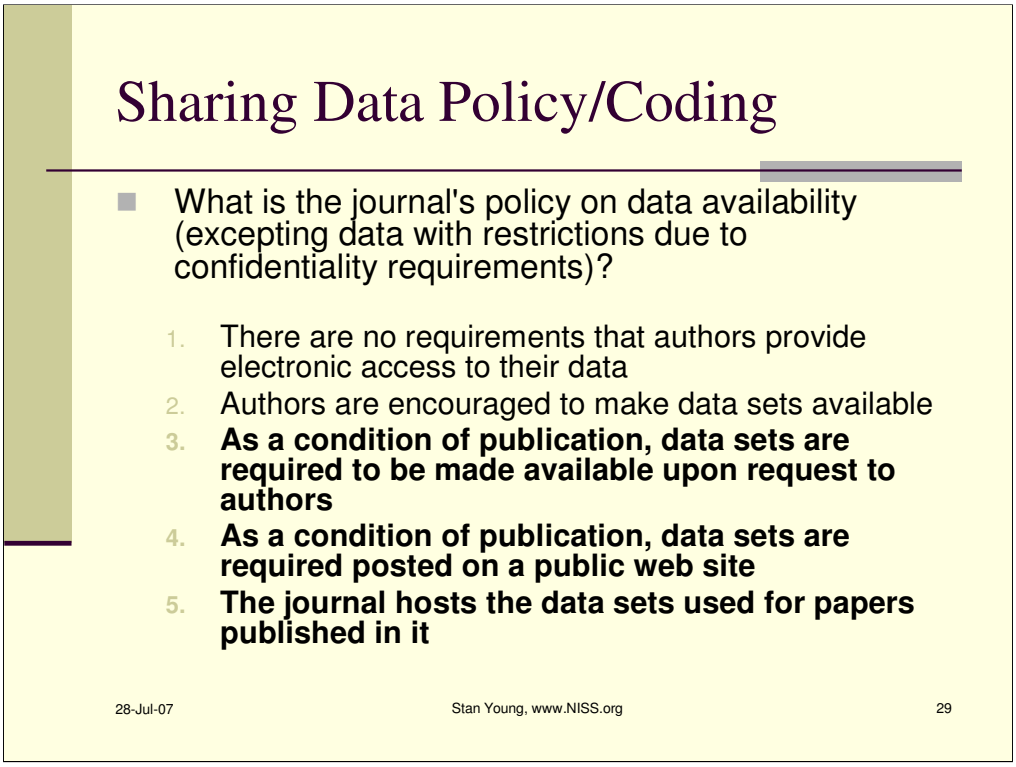
28-Jul-07 Stan Young, www.NISS.org 28

An editor get a 1 for each “correct” answer (multiple testing required and data sharing encouraged).

Epidemiolgists have an average multiple testing score of 0.63; others have an average score of 1.21.

The results were statistically significant as 0.019 < 0.040.

Conclusion: epidemiology editors are less likely to address multiple testing than other science. Somewhat as expected, RCT scientist require adjustment for multiple testing.



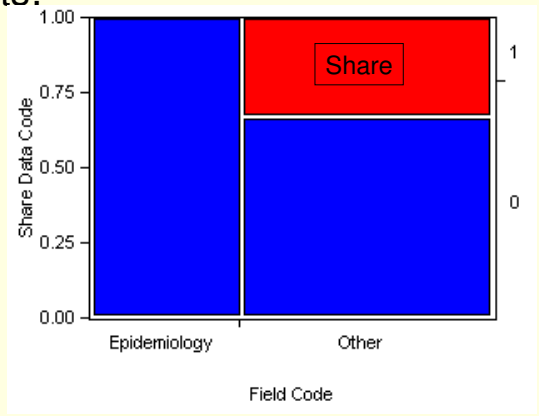
It is important to note that sharing of data and research materials is suppose to be one of the tenets of science. The National Academy studied the question and have a research monograph on the subject:

Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences

http://books.nap.edu/catalog/10613.html

Analysis

- Data Sharing:
  - Epidemiologists: 0/18
  - Other: 10/30
  - P=0.0077
    (Fisher's Exact Test)



28-Jul-07 Stan Young, www.NISS.org 30

18 editors from epidemiology journals say, all 18 in the survey, that there is no policy for data sharing with their journals. Only about 1/3 of the journals encourage data sharing. As a point of reference, some major journals require data sharing.

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

1/ 20, 5% !!

28-Jul-07 Stan Young, www.NISS.org 31

So we return to the problem. The false discovery rate for epidemiolgy is empirically 80 to 90%.

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.

Nature and necessity of scientific revolutions

"…existing institutions have ceased adequately to meet the problems posed by an environment that they have in part created."

"… an existing paradigm has ceased to function adequately in the exploration of an aspect of nature to which that paradigm itself had previously led the way."

The need for a paradigm shift "could be discovered only through something's first going wrong with normal research."

Kuhn, 1962

28-Jul-07 Stan Young, www.NISS.org 32

When is there a crisis?

A crisis occurs when normal science starts making mistakes. For example, randomized clinical trials are not confirming claims coming from non-randomized trial; low fat diets are not confirming claims of lower heart attack rates, fewer strokes, lower rates of colon cancer, breast cancer. Also stress and high blood pressure, type A personality and heart attacks, etc.

An informal count puts the claims of epidemiologists being supported in randomized clinical trials at ~1-2 out of 20.

Crisis?

1. People are noticing problems

Ioannidis (2005): 5/6 observational studies fail to replicate.

2. “Though they may begin to lose faith and then to consider alternatives, they do not renounce the paradigm that has led them into crisis.” (Kuhn)

3. Mostly epidemiologists consider the failure of only one study at a time. They point to everything except multiple testing. (*Two recent exceptions*.)

Most typically, Type one error, false discovery, chance, etc. is not used by epidemiologists as an explanation for failure to replicate.

GSK, Avandia, Meta Analysis

- A meta-analysis of 42 clinical studies (mean duration 6 months; 14,237 total patients), most of which compared AVANDIA to placebo, showed AVANDIA to be associated with an increased risk of myocardial ischemic events such as angina or myocardial infarction. Three other studies (mean duration 41 months; 14,067 patients), comparing AVANDIA to some other approved oral antidiabetic agents or placebo, have not confirmed or excluded this risk. In their entirety, the available data on the risk of myocardial ischemia are inconclusive. (5.2)

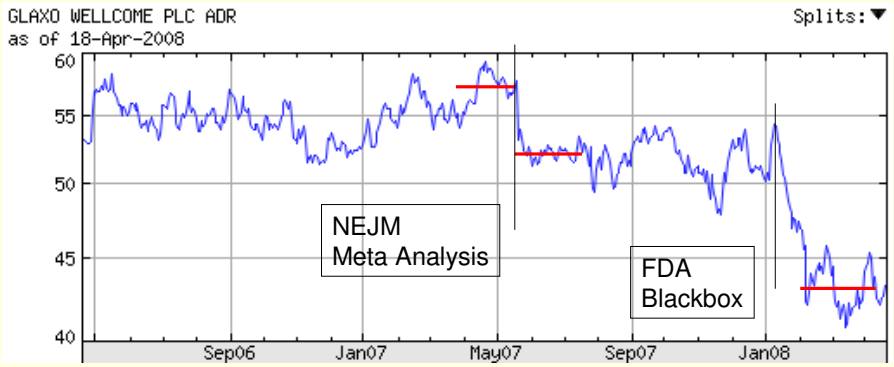
WSJ Report

Contrary to the meta-analysis, WellPoint's initial findings didn't necessarily indicate a higher heart-attack risk linked to Avandia than to Actos and other diabetes medications.

What could well be a flawed meta analysis appeared in the NEJM.

Two large observational studies do not confirm the meta-analysis.

Consequences



GSK lost $38.2B in market cap
and is cutting thousands of jobs.

A share price of 57.5 gives a market cap of 154.6B,
52.0 = 139.8B,
43.3 = 116.4B.

GSK lost $38.2B in market cap.

~20% of sales goes into research. It is estimated to cost $1.8B to bring a drug to market. GSK could have paid for four new drugs with this loss. Arguably, GSK share holders would be richer and society would be better off.

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.

Pocock says that epidemiolgy is in crisis.

Inannidis points out the ~80% false discovery rate of epidemiology.

Rothman (1990) says no correction for multiple testing is necessary and Vandenbroucke, PLoS Med (2008) agrees.

Shapiro will have nothing of the standard epidemiology paradigm 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. His paper is an easy and fun read.