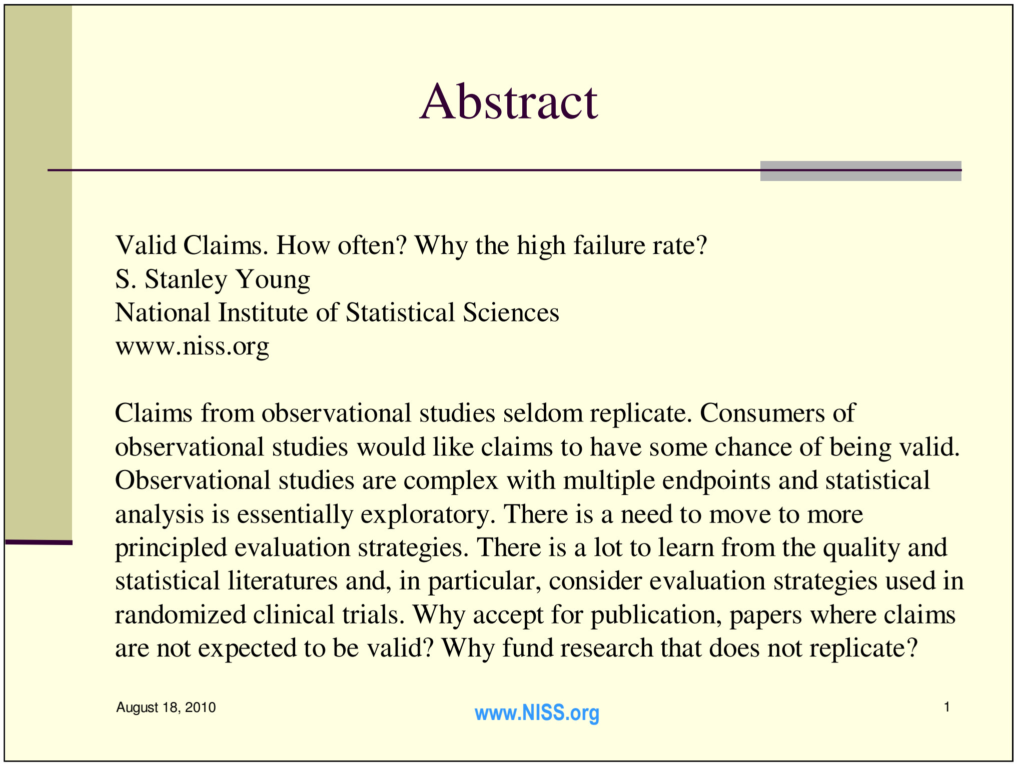


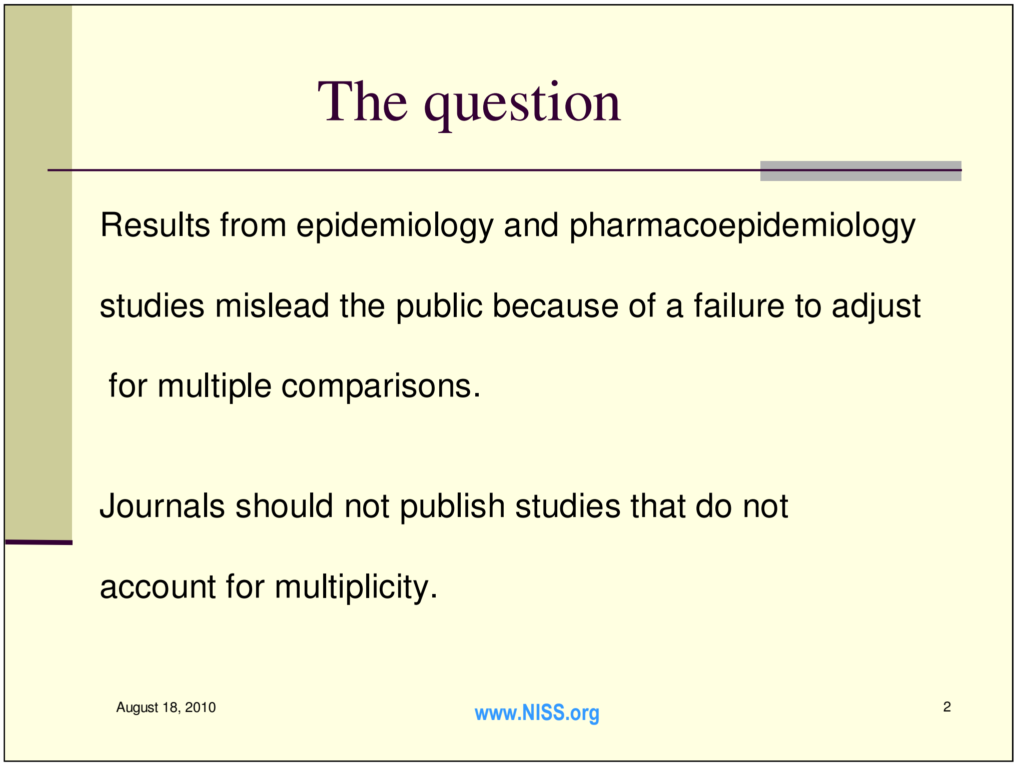
Abstract

Valid Claims. How often? Why the high failure rate?
S. Stanley Young
National Institute of Statistical Sciences
www.niss.org

Claims from observational studies seldom replicate. Consumers of observational studies would like claims to have some chance of being valid. Observational studies are complex with multiple endpoints and statistical analysis is essentially exploratory. There is a need to move to more principled evaluation strategies. There is a lot to learn from the quality and statistical literatures and, in particular, consider evaluation strategies used in randomized clinical trials. Why accept for publication, papers where claims are not expected to be valid? Why fund research that does not replicate?

It is an observed fact that claims coming from well-conducted observational studies most often fail to replicate when tested in randomized clinical trials. Ioannidis, JAMA, 2005.

That claims coming from observation studies most often fail to replicate indicates that it would be well to introduce more rigor into the evaluation strategies of observational studies.



The question

Results from epidemiology and pharmacoepidemiology studies mislead the public because of a failure to adjust for multiple comparisons.

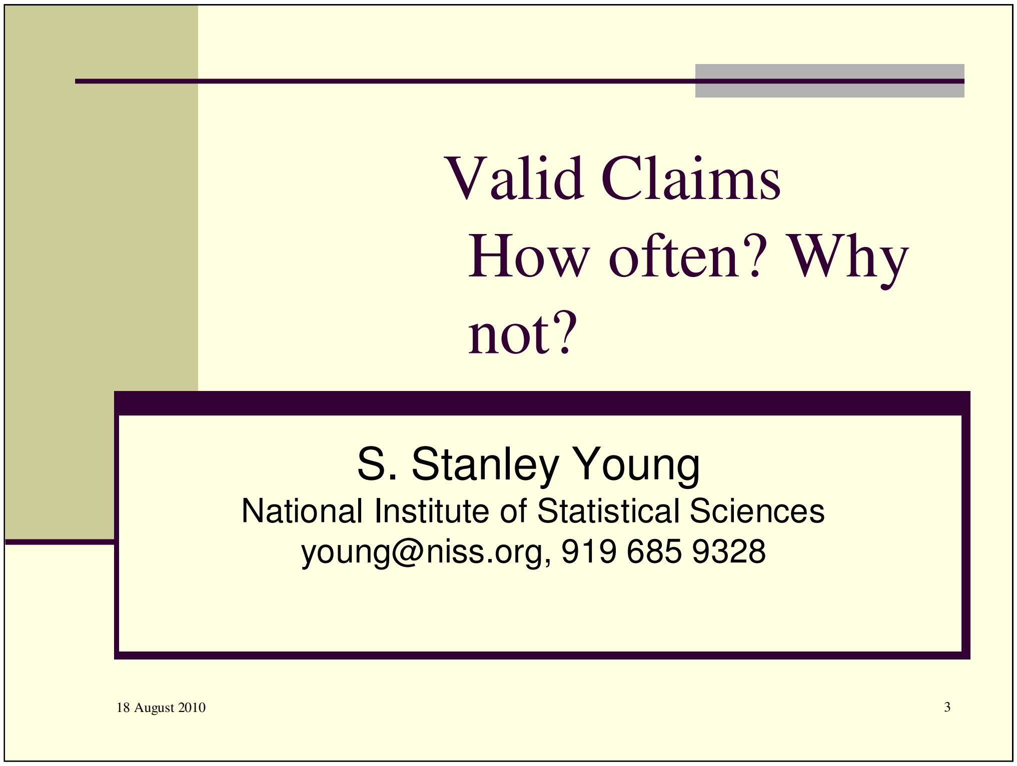
Journals should not publish studies that do not account for multiplicity.

August 18, 2010

www.NISS.org

2

All epidemiology studies need to tell the reader how many questions are at issue. Data sets and analysis code should be publicly available. These studies are most often publicly funded so the whole study process should be as open and transparent at possible. Those conducting these studies should be helping the consumers make good decisions.

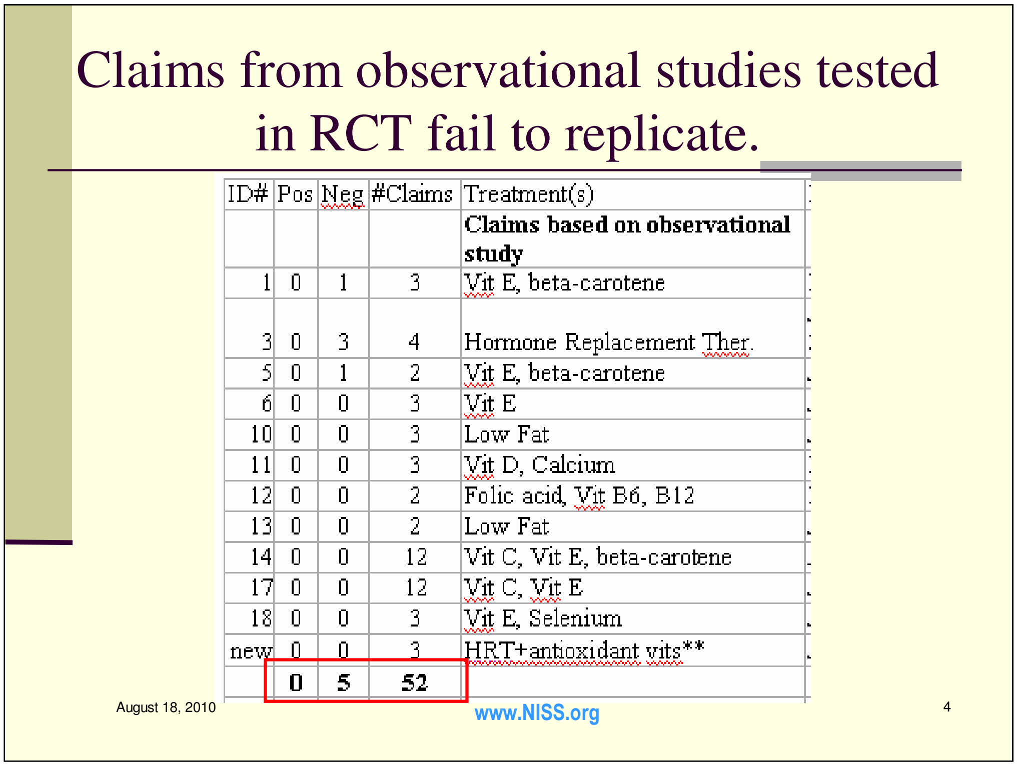


Valid Claims How often? Why not?

S. Stanley Young

National Institute of Statistical Sciences

young@niss.org, 919 685 9328



Claims from observational studies tested in RCT fail to replicate.

ID#	Pos	Neg	#Claims	Treatment(s)
				Claims based on observational study
1	0	1	3	Vit E, beta-carotene
3	0	3	4	Hormone Replacement Ther.
5	0	1	2	Vit E, beta-carotene
6	0	0	3	Vit E
10	0	0	3	Low Fat
11	0	0	3	Vit D, Calcium
12	0	0	2	Folic acid, Vit B6, B12
13	0	0	2	Low Fat
14	0	0	12	Vit C, Vit E, beta-carotene
17	0	0	12	Vit C, Vit E
18	0	0	3	Vit E, Selenium
new	0	0	3	HRT+antioxidant vits**
	0	**5**	**52**	

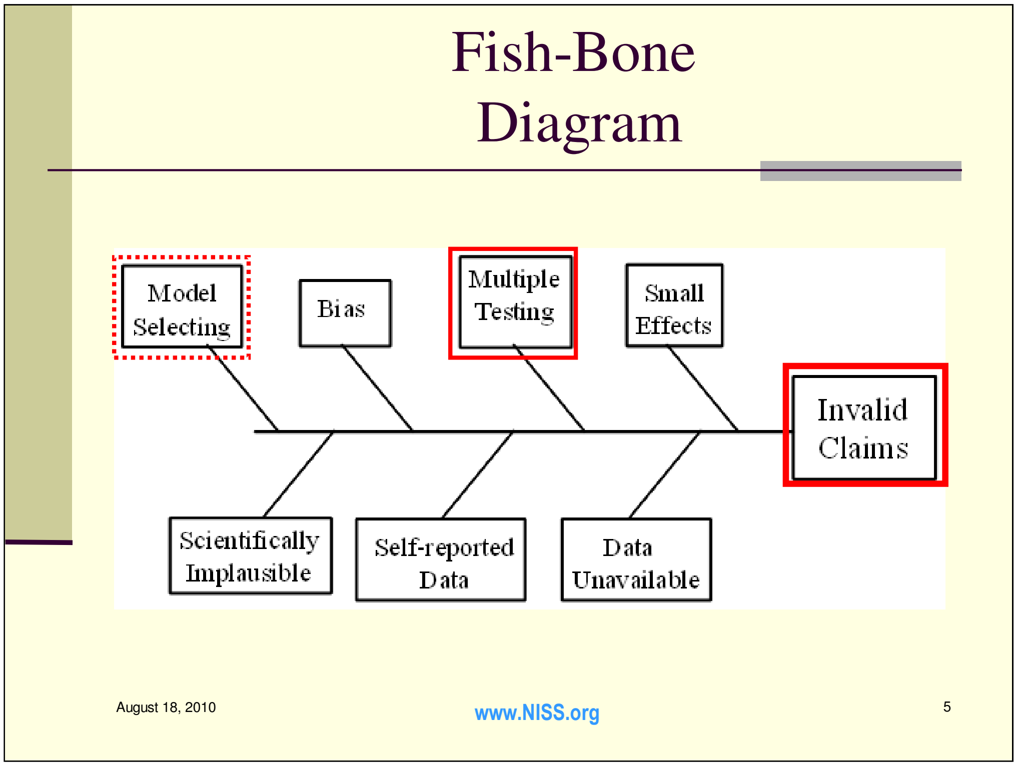
August 18, 2010

www.NISS.org

4

It is an observed fact that claims coming from well-conducted observational studies most often fail to replicate when tested in randomized clinical trials. Ioannidis, JAMA, 2005.

That claims coming from observation studies most often fail to replicate indicates that it would be well to introduce more rigor into the evaluation strategies of observational studies.



Fish-Bone Diagram

Model Selecting

Bias

Multiple Testing

Small Effects

Invalid Claims

Scientifically Implausible

Self-reported Data

Data Unavailable

August 18, 2010

www.NISS.org

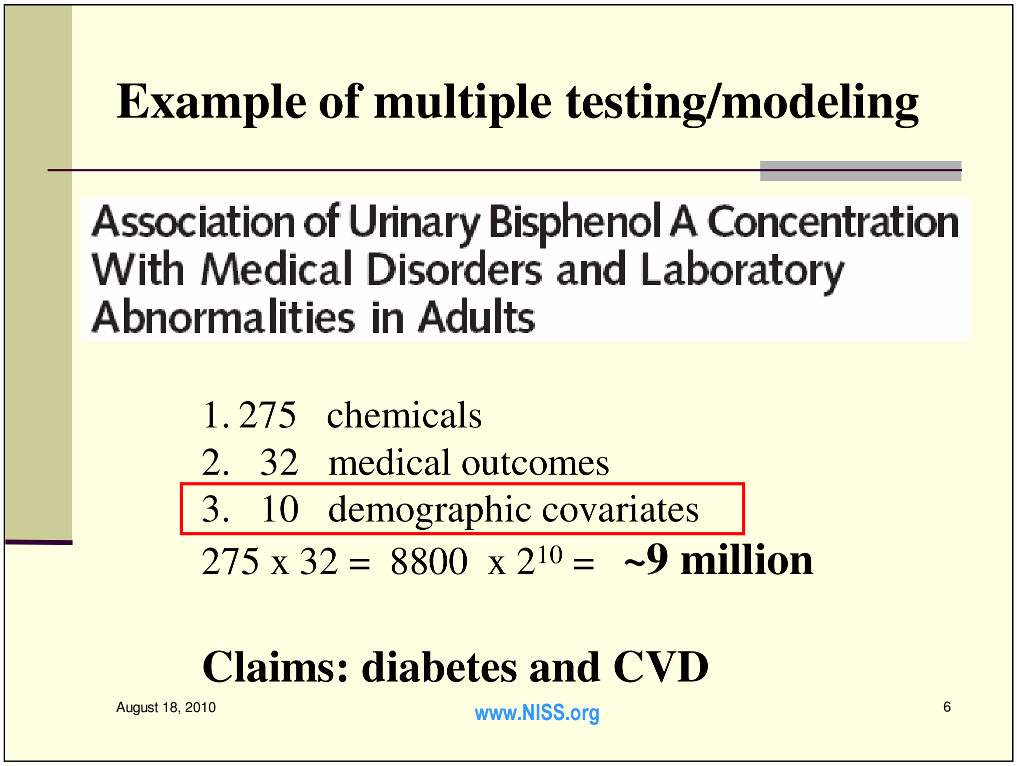
5

What are the possible reasons for invalid claims coming from observational studies? The classic “cause and effect” diagram can be used to put the possible explanatory factors in front of us.

Multiple testing is obvious. It is well-studied and there is good technology for a solution. See Westfall and Young, Wiley, 1993.

Multiple modeling and how that affects false discovery rate is less well studied.

Using one half of the data to generate hypotheses and the other to confirm them or not is a widely used, standard way to deal with multiple modeling and multiple testing. Epidemiologists rarely use the split sample method.



Example of multiple testing/modeling

Association of Urinary Bisphenol A Concentration With Medical Disorders and Laboratory Abnormalities in Adults

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

275 x 32 = 8800 x 2^{10} = **~9 million**

Claims: diabetes and CVD

August 18, 2010

www.NISS.org

6

It is relatively easy to generate 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. (Likewise it would be a tragedy to unnecessarily limit patient options for HIV.)

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.

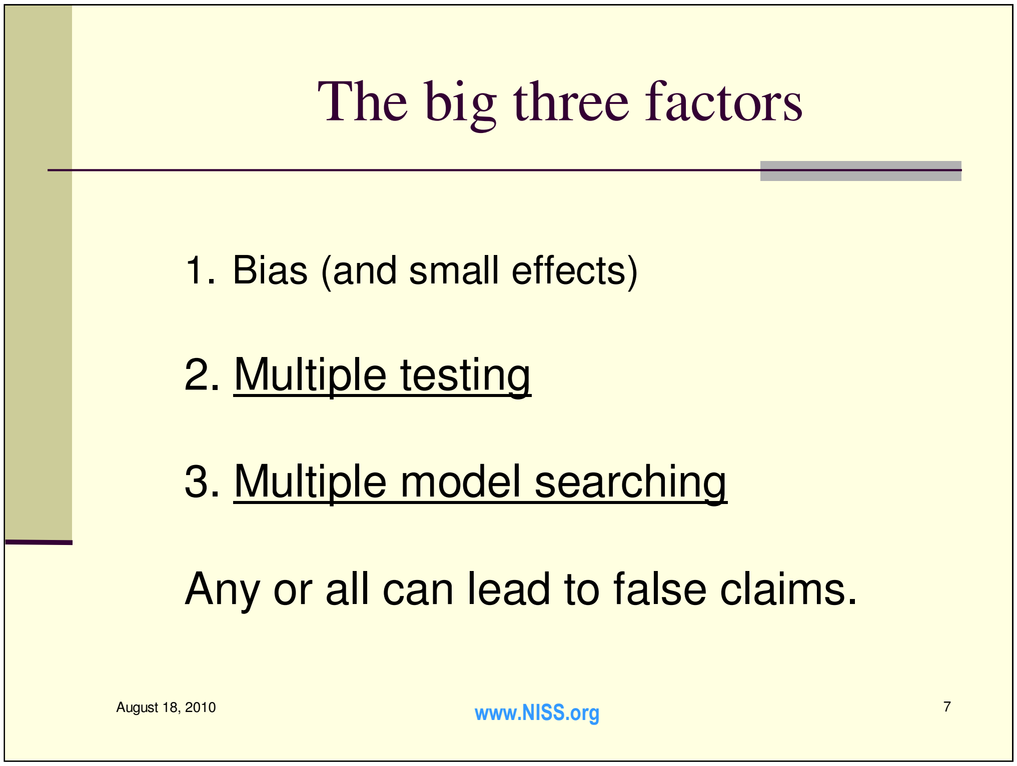
The CDC NHANES (2003-2004) measured 275 environmental chemicals and a wide range of health outcomes. Although the study by Lang et al focused on 1 chemical and 16 health outcomes (8 patient-reported medical outcomes and 8 clinical chemistry measurements), counting to determine how many questions were 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 32275=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 for managers to step back and consider the entire CDC study for the large, planned study that it is and develop an analysis strategy that takes into account the large number of questions at issue



The big three factors

1. Bias (and small effects)

2. Multiple testing

3. Multiple model searching

Any or all can lead to false claims.

August 18, 2010

www.NISS.org

7

Unless the statistical analysis of observational studies is carefully done, every study will have one or more positive effects. All of the claims could well be false positives.

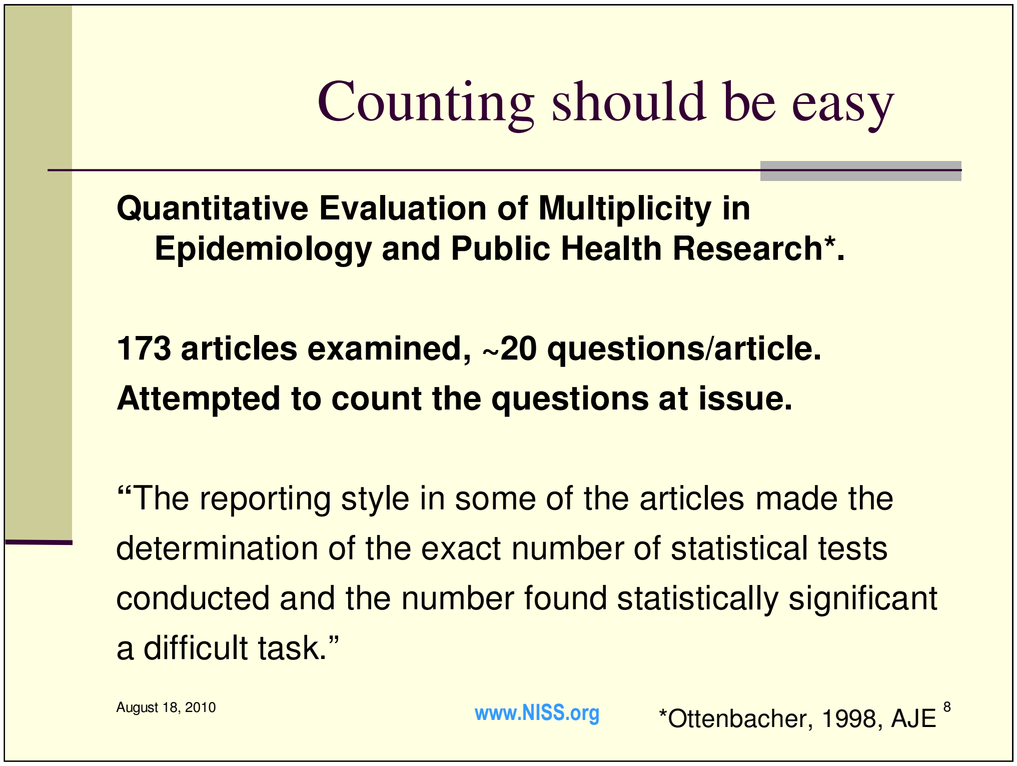
The word bias covers a lot of sins. Unmeasured confounders. Measured, but unused confounders. Modeling bias – run hundreds of models and select the one you like.

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 option 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. Editors reward p-value searching and punish “no effect” studies. Editors are part of the problem.



Counting should be easy

Quantitative Evaluation of Multiplicity in Epidemiology and Public Health Research*.

173 articles examined, ~20 questions/article.

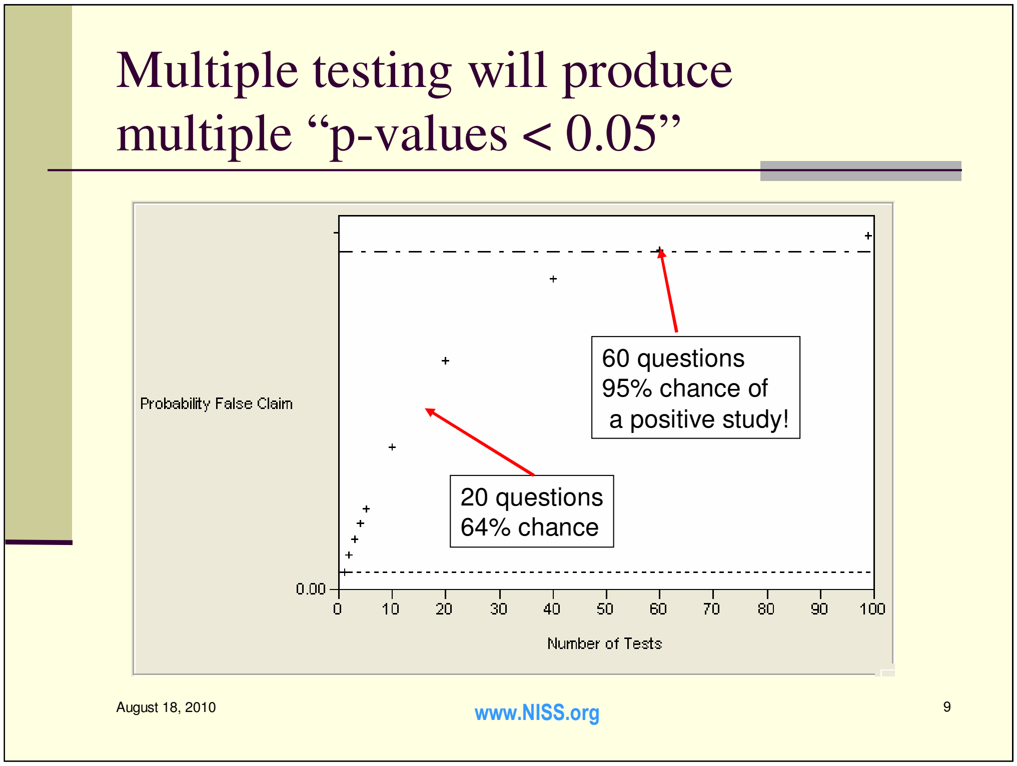
Attempted to count the questions at issue.

"The reporting style in some of the articles made the determination of the exact number of statistical tests conducted and the number found statistically significant a difficult task."

August 18, 2010 www.NISS.org *Ottenbacher, 1998, AJE 8

It is easy to progress with an exploratory analysis and completely mess up any chance of counting the questions at issue. It makes sense to plan the analysis and think about what makes sense before starting the analysis.

Authors can and do obfuscate.



Multiple testing will produce multiple “p-values < 0.05”

60 questions
95% chance of
a positive study!

20 questions
64% chance

August 18, 2010

www.NISS.org

9

Ask a lot of questions in a study and you are very likely to get statistically significant results by chance alone.

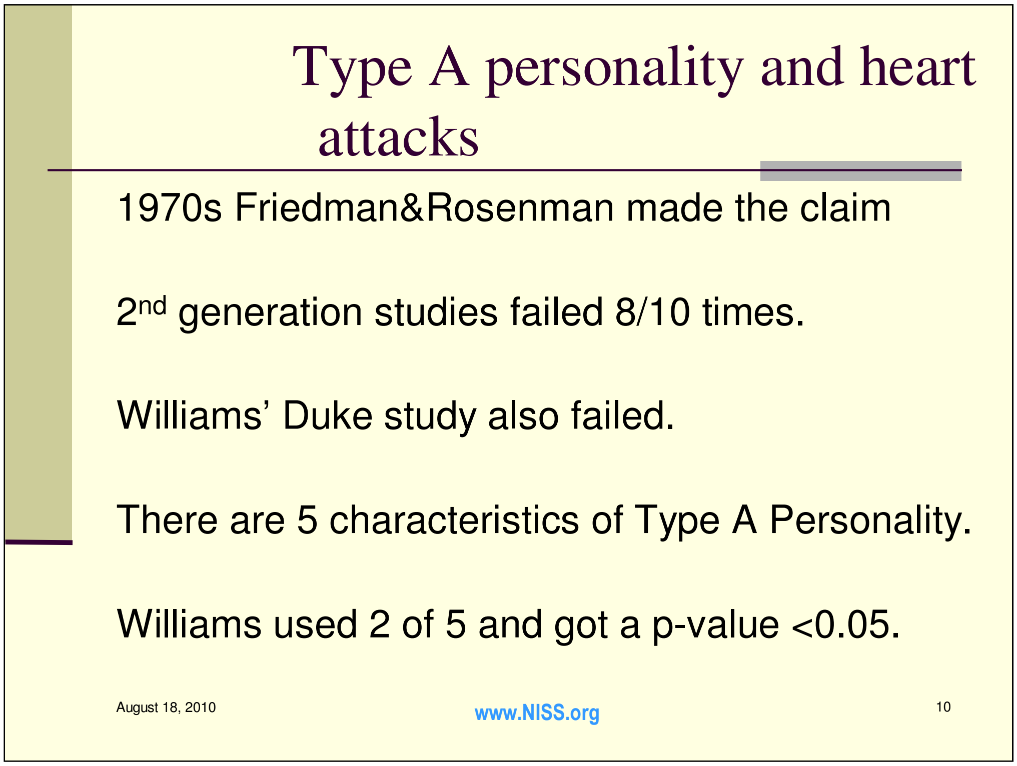
If 60 independent questions are asked in an experiment there is a 95% probability of at least one “statistically significant” result.

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 the first task is to carefully specify the questions at issue and how data and analysis will be used to evaluate those questions.

It is fairly typical in observational studies to ask a lot of questions and not clearly state how many questions are at issue.

In a RCT, the number of questions at issue is explicitly given as part of the protocol.



Type A personality and heart attacks

1970s Friedman&Rosenman made the claim

2nd generation studies failed 8/10 times.

Williams' Duke study also failed.

There are 5 characteristics of Type A Personality.

Williams used 2 of 5 and got a p-value <0.05.

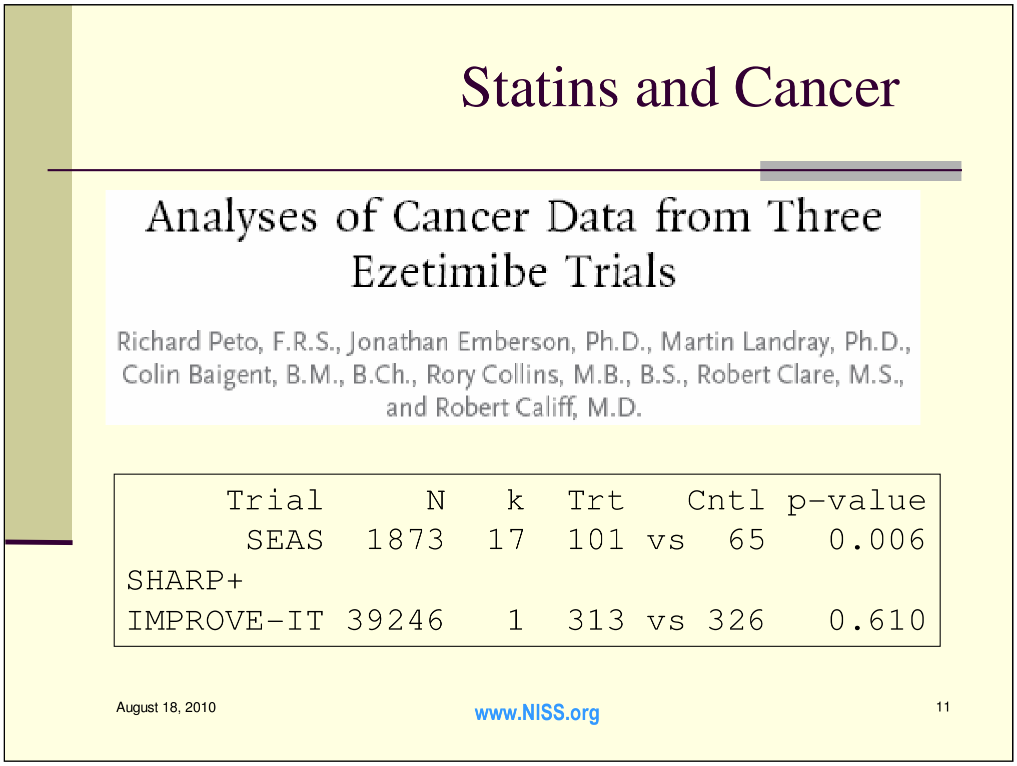
August 18, 2010 www.NISS.org 10

There are 31 possible models with 5 predictors.

Multiple modeling and retrospective rationalization appear to be in play.

Psychologists no longer consider Type A personality an official personality type.

The myth lives on.



Statins and Cancer

Analyses of Cancer Data from Three Ezetimibe Trials

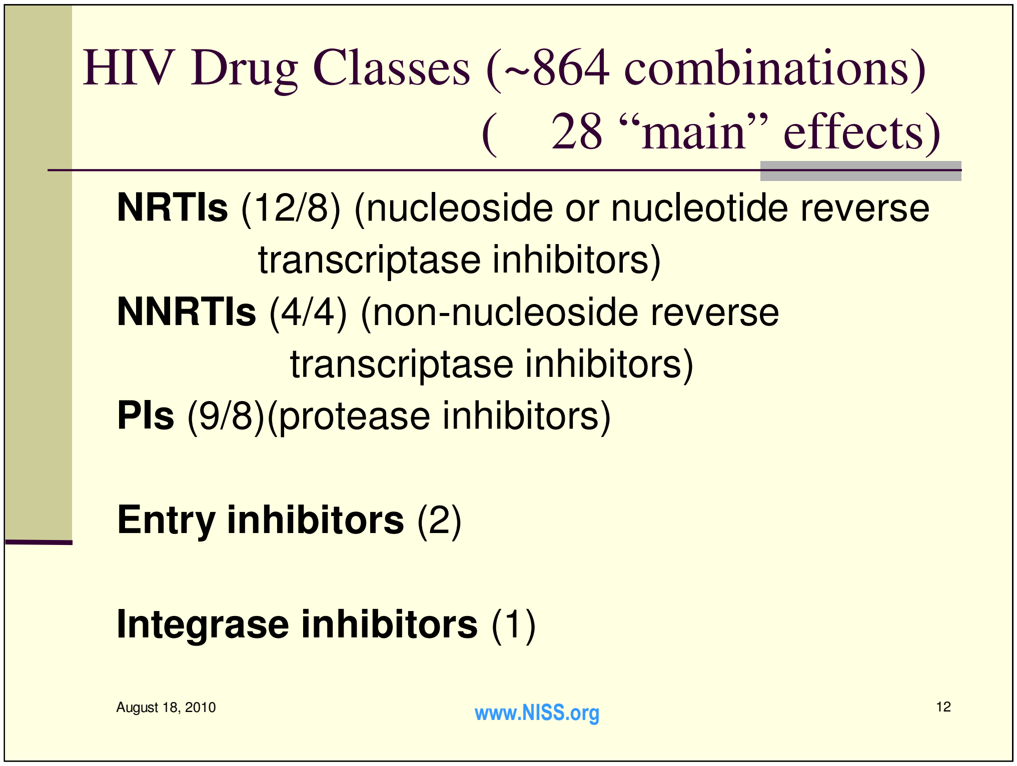
Richard Peto, F.R.S., Jonathan Emberson, Ph.D., Martin Landray, Ph.D., Colin Baigent, B.M., B.Ch., Rory Collins, M.B., B.S., Robert Clare, M.S., and Robert Califf, M.D.

Trial	N	k	Trt		Cntl	p-value
SEAS	1873	17	101	vs	65	0.006
SHARP+ IMPROVE-IT	39246	1	313	vs	326	0.610

Analysis of safety variables in a RCT gave rise to the claim that statins cause cancer.

There were at least 17 tests of hypothesis in the SEAS trial.

The validity of the claim was not supported in two follow on RCTs where one claim was tested. The first claim was a false positive. There was regression to the mean.



HIV Drug Classes (~864 combinations) (28 “main” effects)

NRTIs (12/8) (nucleoside or nucleotide reverse transcriptase inhibitors)
NNRTIs (4/4) (non-nucleoside reverse transcriptase inhibitors)
PIs (9/8)(protease inhibitors)

Entry inhibitors (2)

Integrase inhibitors (1)

August 18, 2010 www.NISS.org 12

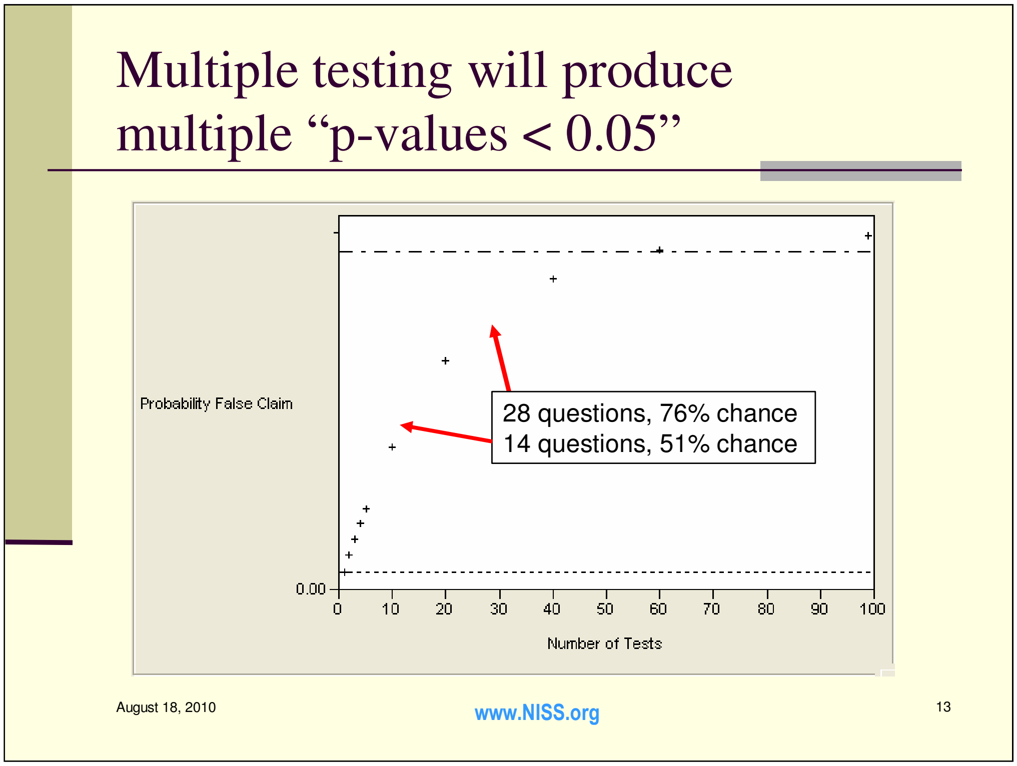
There are a large number of treatments under consideration. I assume that only a relatively few of the combinations are used extensively.

There are many ways to slice and dice the data. Look at all 28. Look at the five classes. Look at the drugs within each class.

There is also much treatment switching.

So far in my reading there is no clear support for a biological mechanism for CV effects.

HIV drugs are unlikely to cause MIs.



Multiple testing will produce multiple “p-values < 0.05”

28 questions, 76% chance
14 questions, 51% chance

August 18, 2010

www.NISS.org

13

Ask a lot of questions in a study and you are very likely to get statistically significant results by chance alone.

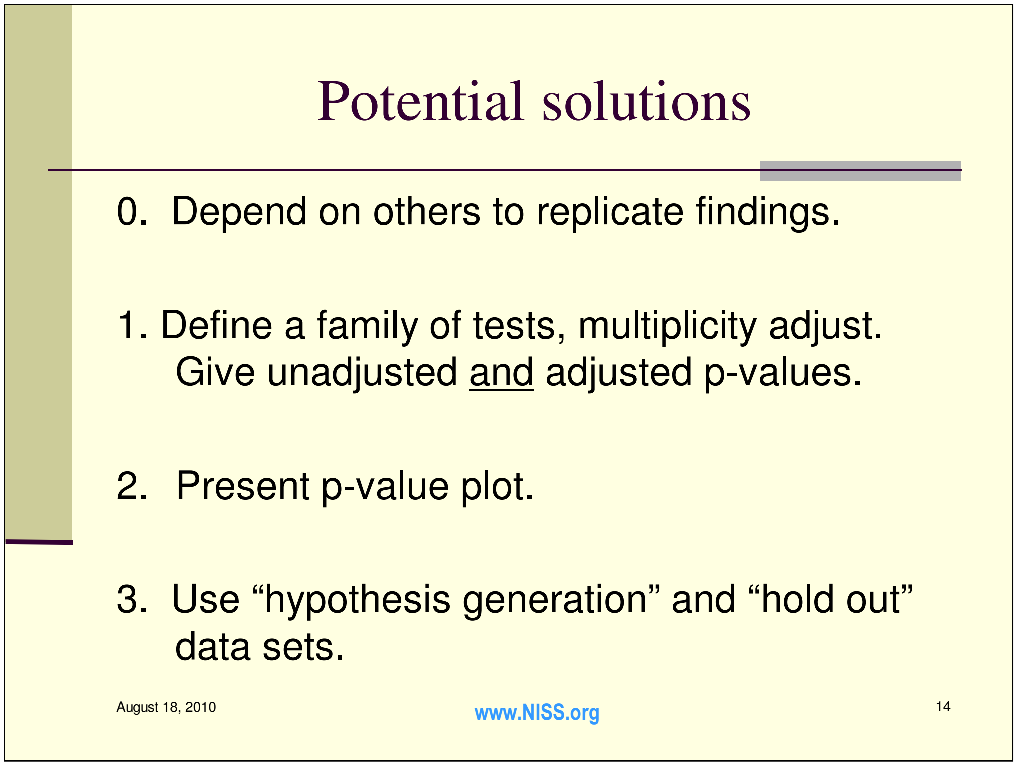
If 14 independent questions are asked in an experiment there is a 51% probability of at least one “statistically significant” result. With 28 tests, a 76% chance of a false positive.

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 the first task is to carefully specify the questions at issue and how data and analysis will be used to evaluate those questions.

It is fairly typical in observational studies to ask a lot of questions and not clearly state how many questions are at issue.

In a RCT, the number of questions at issue is explicitly given as part of the protocol and is most typically very small, 1-3 or so.



Potential solutions

0. Depend on others to replicate findings.

1. Define a family of tests, multiplicity adjust.
   Give unadjusted and adjusted p-values.

2. Present p-value plot.

3. Use “hypothesis generation” and “hold out” data sets.

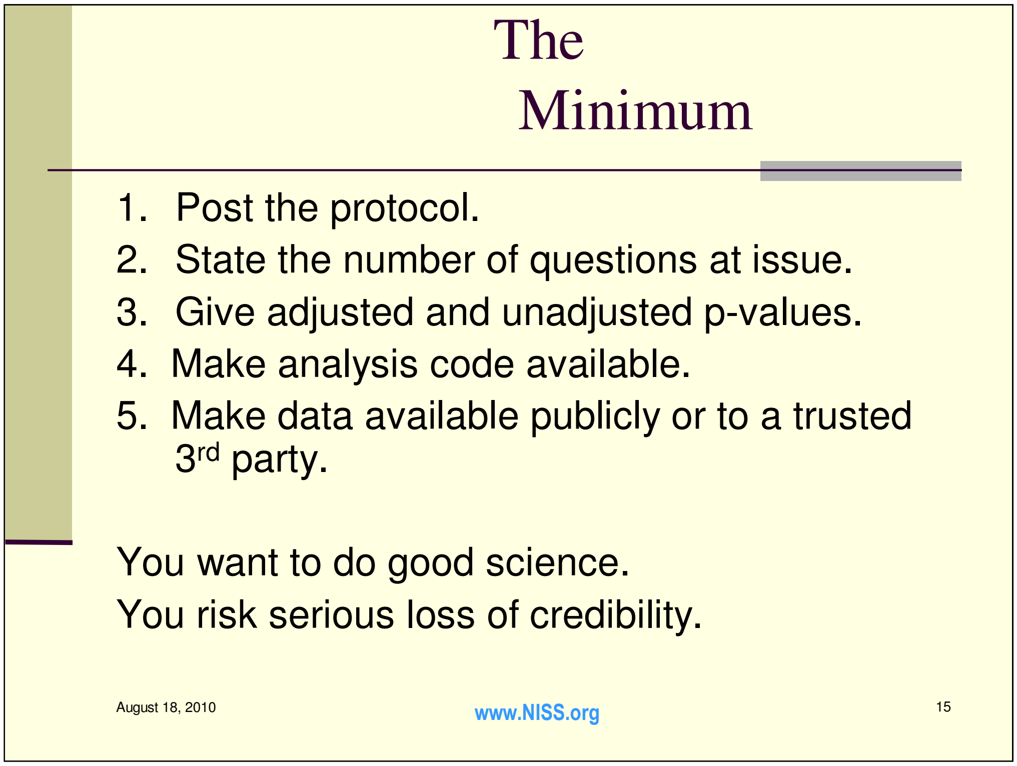
August 18, 2010 www.NISS.org 14

It seems essential that resources be committed to replication in some sense.

The author needs to be fair to the reader. Giving only unadjusted p-values is very unfair.

Adjusted and unadjusted analysis should be done and presented so that the reader can judge the validity of any claims.

To recommend “depend on others” is to not care about valid claims.



The Minimum

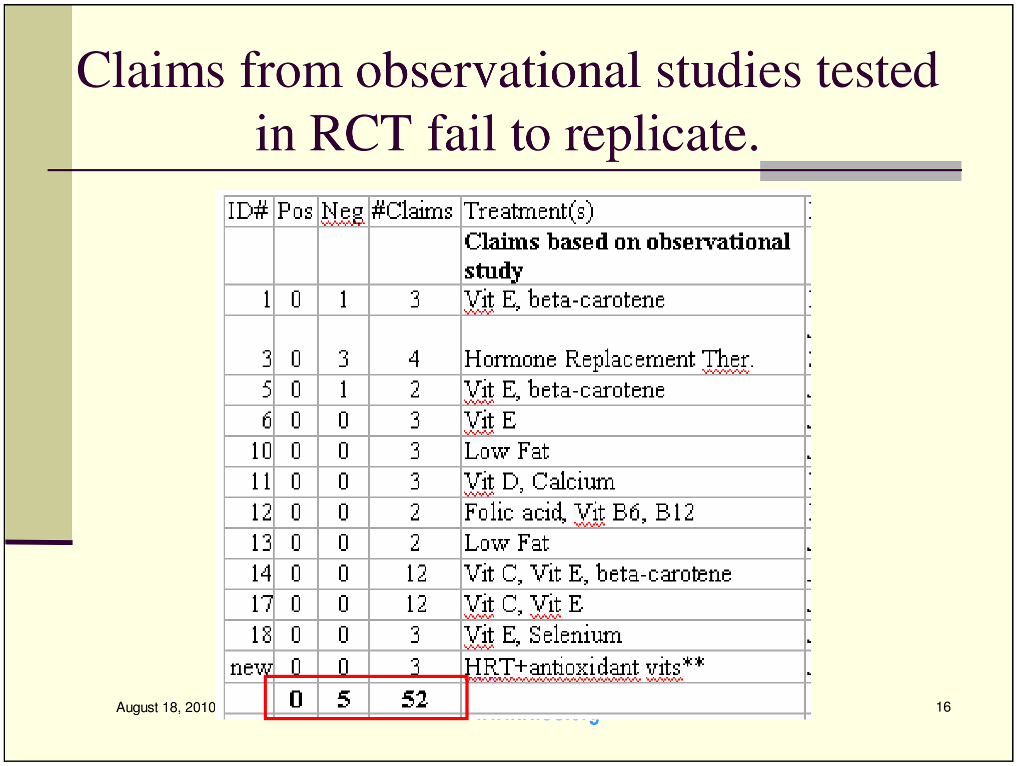
1. Post the protocol.
2. State the number of questions at issue.
3. Give adjusted and unadjusted p-values.
4. Make analysis code available.
5. Make data available publicly or to a trusted 3rd party.

You want to do good science.

You risk serious loss of credibility.

August 18, 2010 www.NISS.org 15

Epidemiolgists want to do good science. They want their claims to be likely to be valid. You risk a serious loss of credibility and loss of funding unless you identify and fix analysis problems.

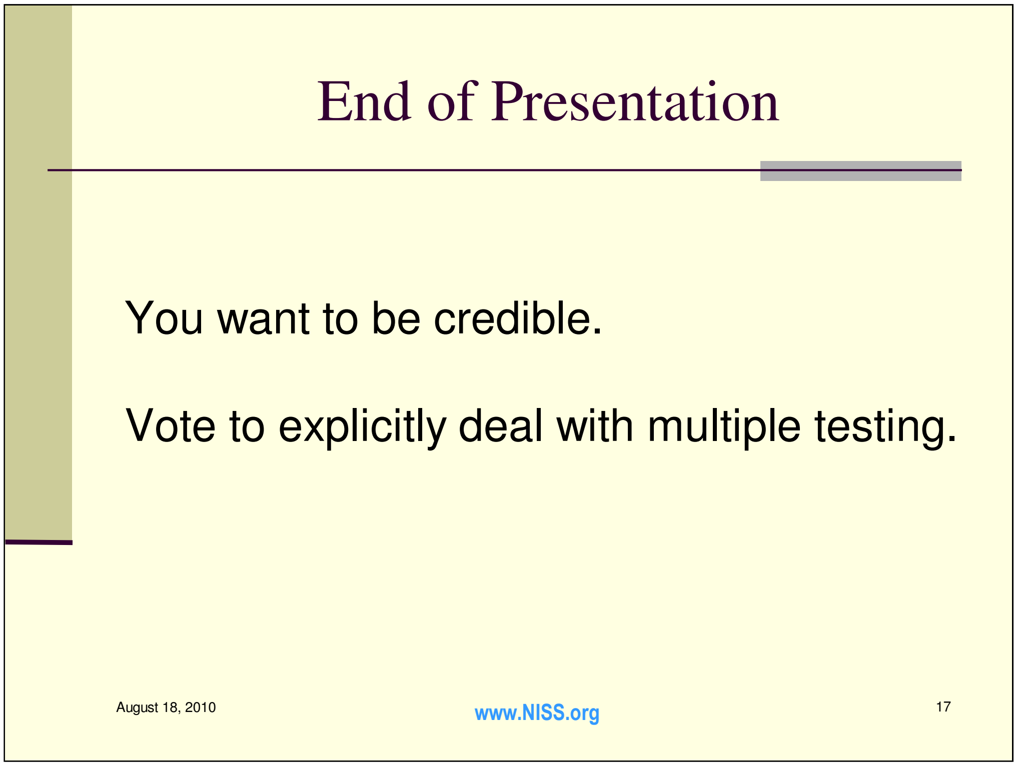


Claims from observational studies tested in RCT fail to replicate.

ID#	Pos	Neg	#Claims	Treatment(s)
				Claims based on observational study
1	0	1	3	Vit E, beta-carotene
3	0	3	4	Hormone Replacement Ther.
5	0	1	2	Vit E, beta-carotene
6	0	0	3	Vit E
10	0	0	3	Low Fat
11	0	0	3	Vit D, Calcium
12	0	0	2	Folic acid, Vit B6, B12
13	0	0	2	Low Fat
14	0	0	12	Vit C, Vit E, beta-carotene
17	0	0	12	Vit C, Vit E
18	0	0	3	Vit E, Selenium
new	0	0	3	HRT+antioxidant vits**
	0	**5**	**52**	

IF epidemiology was producing a string of valid claims, then business as usual would make sense.

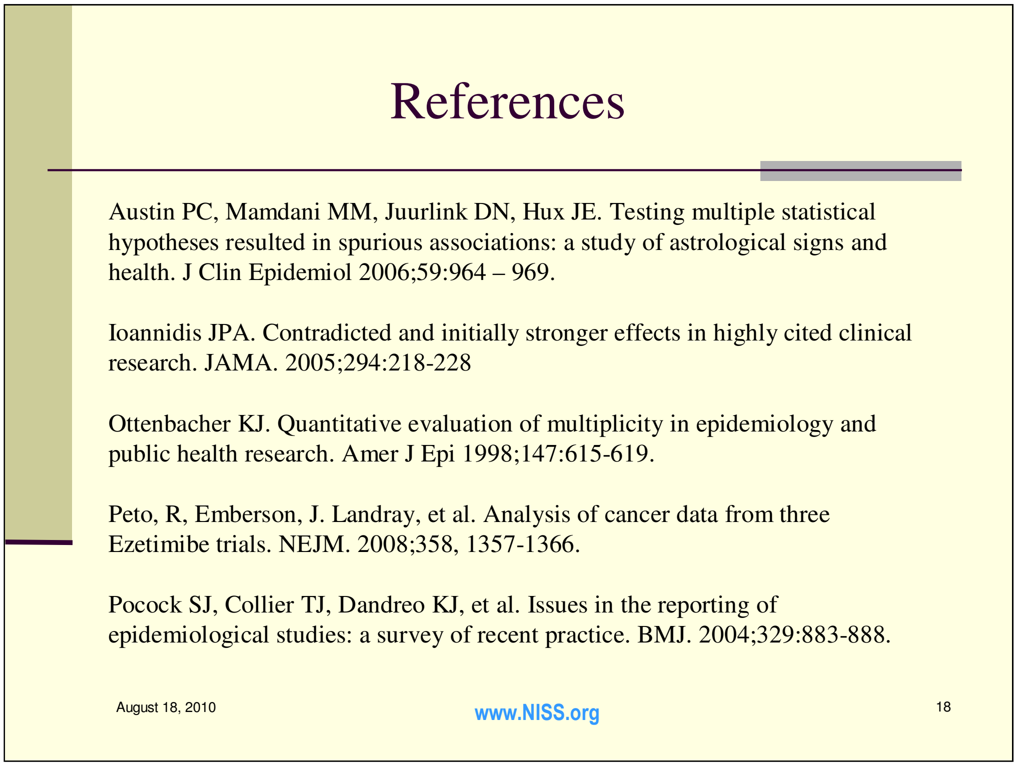
It is the exception when a claim is replicated.Edwards Deming says that when there is a crisis, you can not ask the workers to fix it. It is the job of managers, funding agencies and journal editors, to re-design the system.



End of Presentation

You want to be credible.

Vote to explicitly deal with multiple testing.



References

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.

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

Ottenbacher KJ. Quantitative evaluation of multiplicity in epidemiology and public health research. Amer J Epi 1998;147:615-619.

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

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.

August 18, 2010 www.NISS.org 18

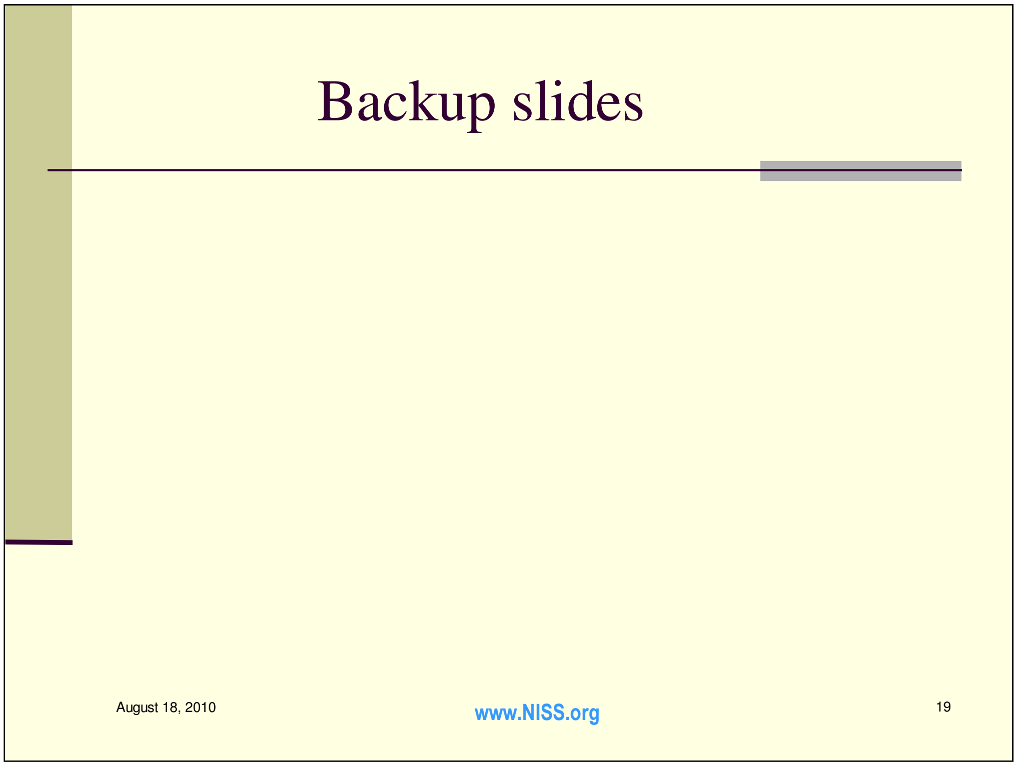
Pocock said that epidemiolgy is in crisis in 2004. The epidemiologists appear (continue to be) to be in denial. Feinstein, 1988, noted the crisis in 1988 and pointed to L. Gardis, 1979, as also pointing to problems.

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

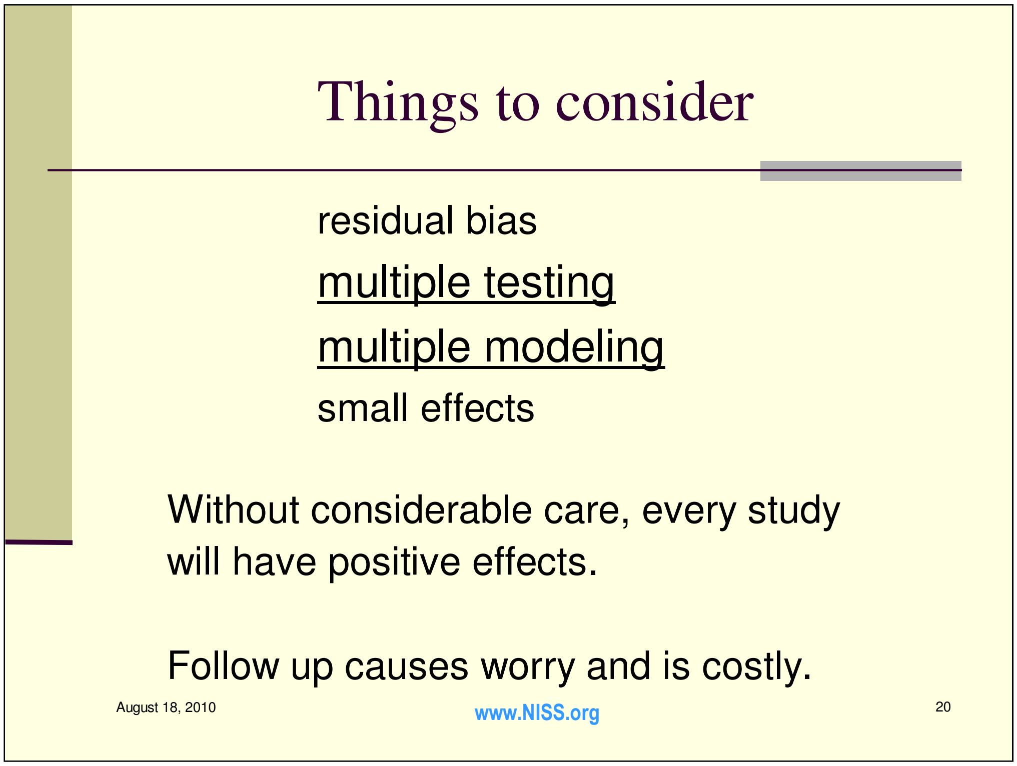
Rothman (1990) and Vandenbrocke (2008) says no correction for multiple testing is necessary. This is the current epidemiolgy paradigm. Basically the workers are happy with the current state of affairs.

Shapiro will have nothing of it and points to an example of a false positive result of 30 years ago from which most epidemiologists seemed to learn nothing. See also Feinstein, Science, 1988.

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



Backup slides



Things to consider

residual bias

multiple testing

multiple modeling

small effects

Without considerable care, every study will have positive effects.

Follow up causes worry and is costly.

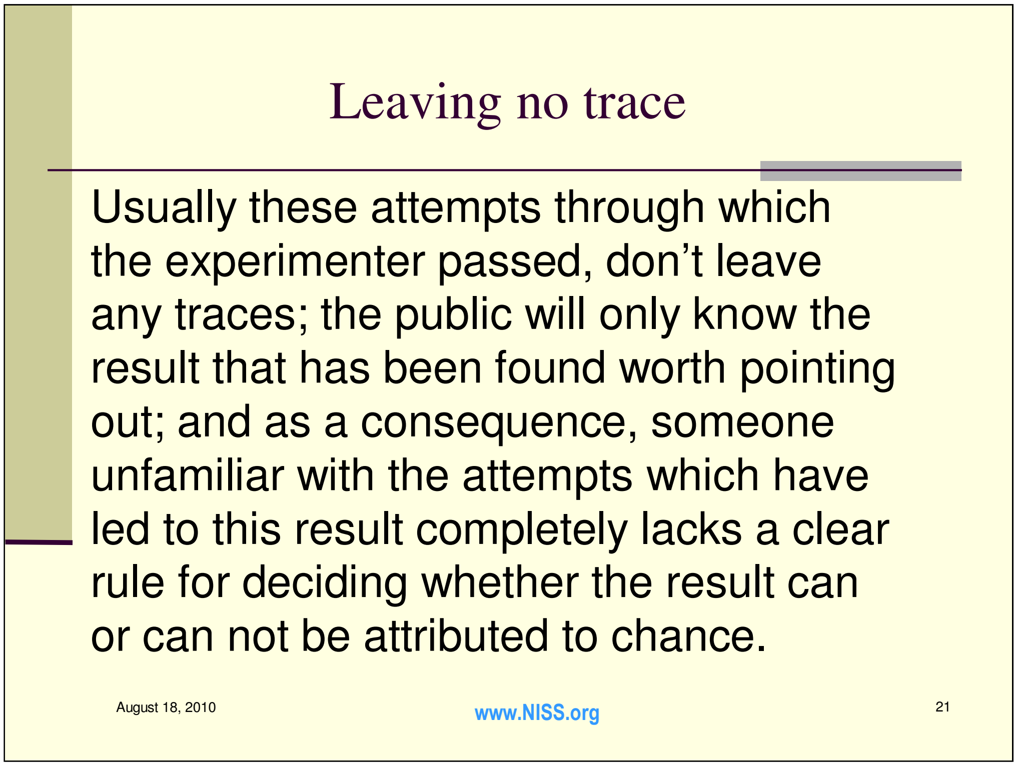
Indiscriminant multiple testing and/or residual bias (and large data sets) can lead to essentially every study having one or more significant effects.

Carefully specify analysis protocol.

Test and holdout data sets.

Specific analysis methods that control multiple testing and multiple model building.

Very few epi claims are subjected to retest. Essentially no epi studies use the simple strategy of dividing the data set in half, then generate hypotheses in one half and test the few questions in the second half. Splitting a data set is a very common procedure for validation of claims.

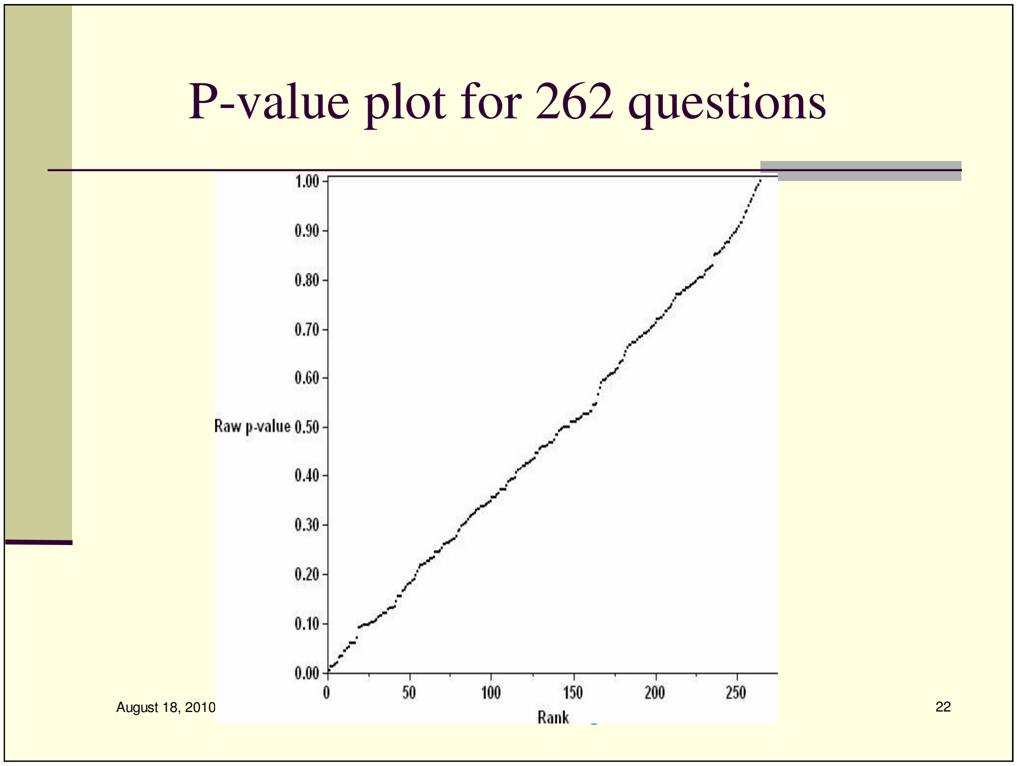


Leaving no trace

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.

Quite important. The epidemiologists, in effect, assume that every question is independently reportable. 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. Most typically, they do not make their data sets public. It is trust me science.

Juliet Shaffer, UC Berkeley



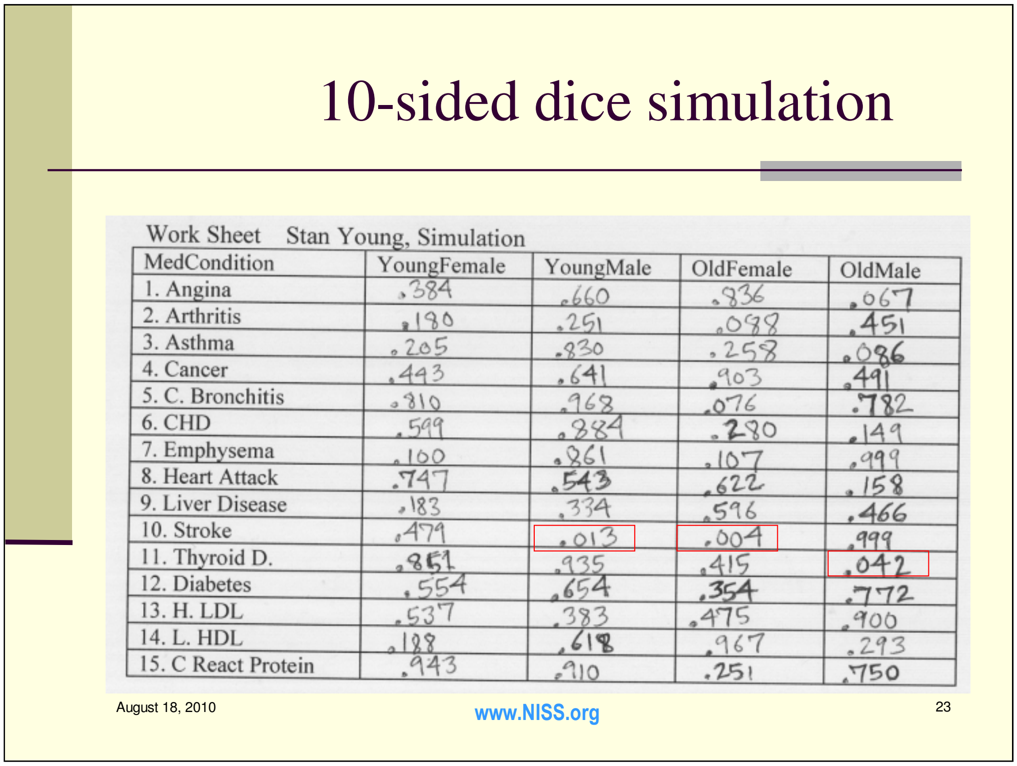
P-value plot for 262 questions

August 18, 2010

22

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). 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.

We conclude that there is no evidence for any nutritional effect on gender, not withstanding the elaborate explanation of the authors and the few small p-values. Adjusted for multiple testing there is no effect.



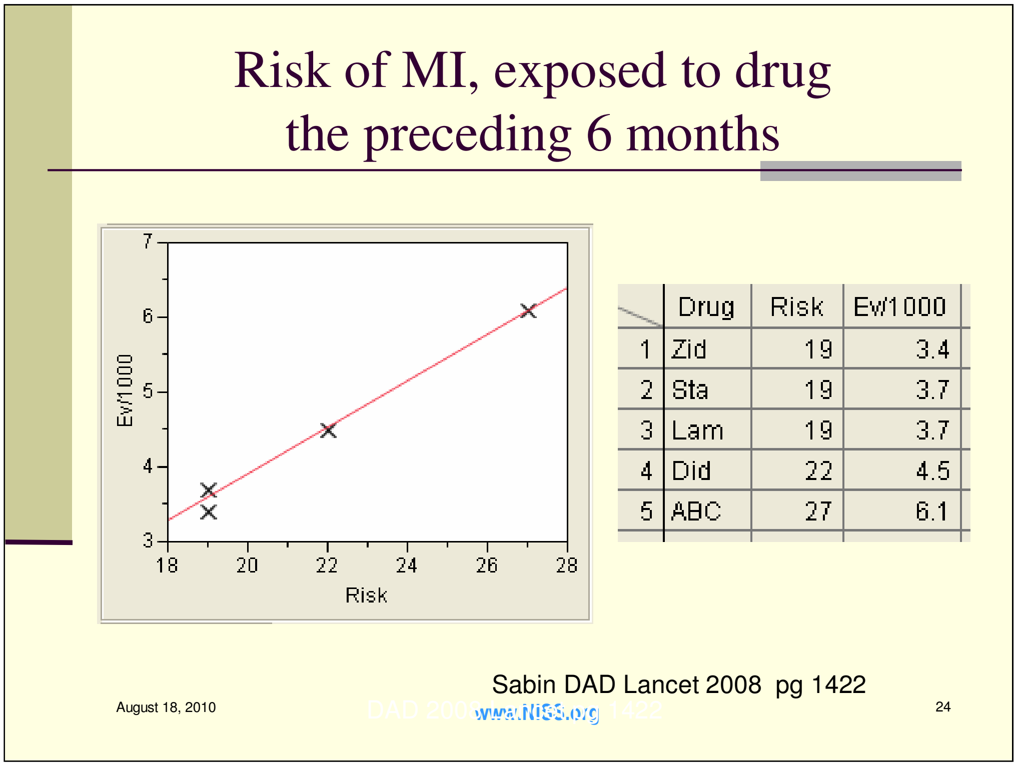
10-sided dice simulation

Work Sheet   Stan Young, Simulation

MedCondition	YoungFemale	YoungMale	OldFemale	OldMale
1. Angina	.384	.660	.836	.067
2. Arthritis	.180	.251	.098	.451
3. Asthma	.205	.830	.258	.086
4. Cancer	.443	.641	.903	.491
5. C. Bronchitis	.810	.968	.076	.782
6. CHD	.599	.884	.280	.149
7. Emphysema	.100	.861	.107	.999
8. Heart Attack	.747	.543	.622	.158
9. Liver Disease	.183	.334	.596	.466
10. Stroke	.479	.013	.004	.999
11. Thyroid D.	.851	.935	.415	.042
12. Diabetes	.554	.654	.354	.772
13. H. LDL	.537	.383	.475	.900
14. L. HDL	.188	.618	.967	.293
15. C React Protein	.943	.910	.251	.750

August 18, 2010   www.NISS.org   23

Three 10-sided dice were rolled 60 times. The simulated p-values were recorded. Probability calculations indicate there should be a 95% chance of at least one statistically significant result. In this simulated experiment, there are three statistically significant results. There is a disconnect here. The person doing this experiment knows the results are random as the person rolled the dice. It is typical to act as if any statistically significant result is operationally real. NB the smallest p-value of 0.004.



Risk of MI, exposed to drug the preceding 6 months

	Drug	Risk	Ev/1000
1	Zid	19	3.4
2	Sta	19	3.7
3	Lam	19	3.7
4	Did	22	4.5
5	ABC	27	6.1

Sabin DAD Lancet 2008 pg 1422

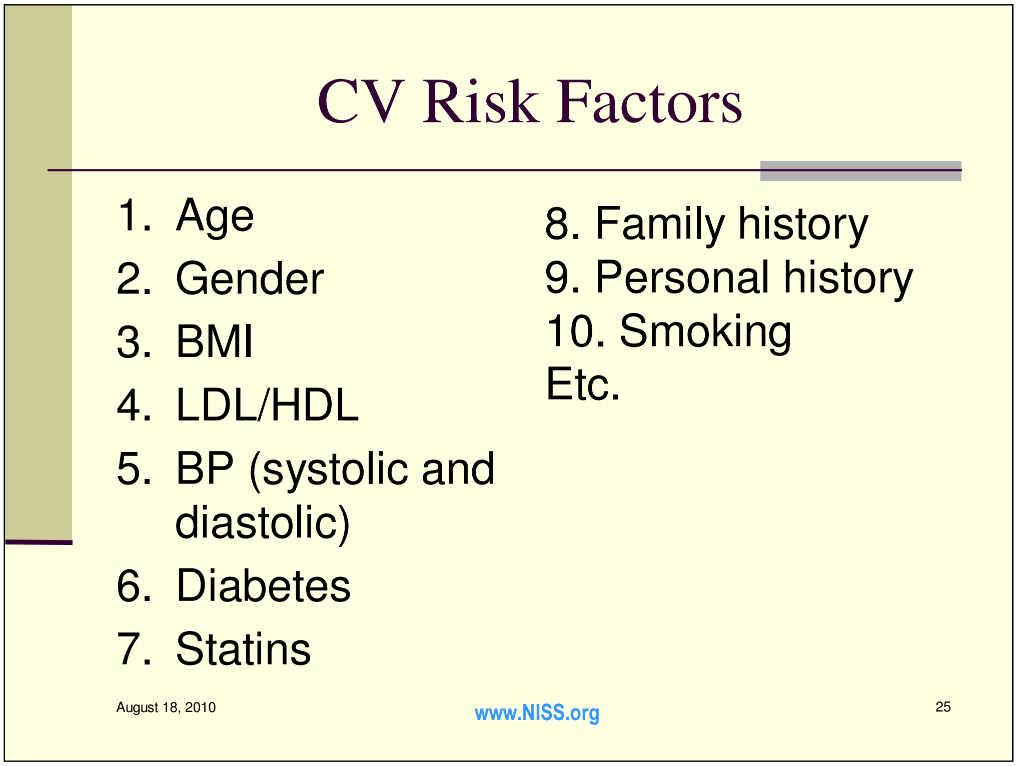
From the Summary:

There exists an increased risk of myocardial infarction in patients exposed to abacavir and didanosine within the preceding 6 months.

Sabin DAD pg 1422

Patients who started abacavir or didanosine for the first time while under follow-up in D:A:D generally had worse cardiovascular risk profiles than did those who started the other NRTIs for the first time: 1119 (27%) of

4076 patients first starting abacavir and 383 (22%) of 1731 first starting didanosine had moderate or high predicted 10-year risk of coronary heart disease compared with 414 (19%) of 2177, 139 (19%) of 741, and 474 (19%) of 2546 patients first starting zidovudine, stavudine, and lamivudine, respectively.



CV Risk Factors

1. Age
2. Gender
3. BMI
4. LDL/HDL
5. BP (systolic and diastolic)
6. Diabetes
7. Statins

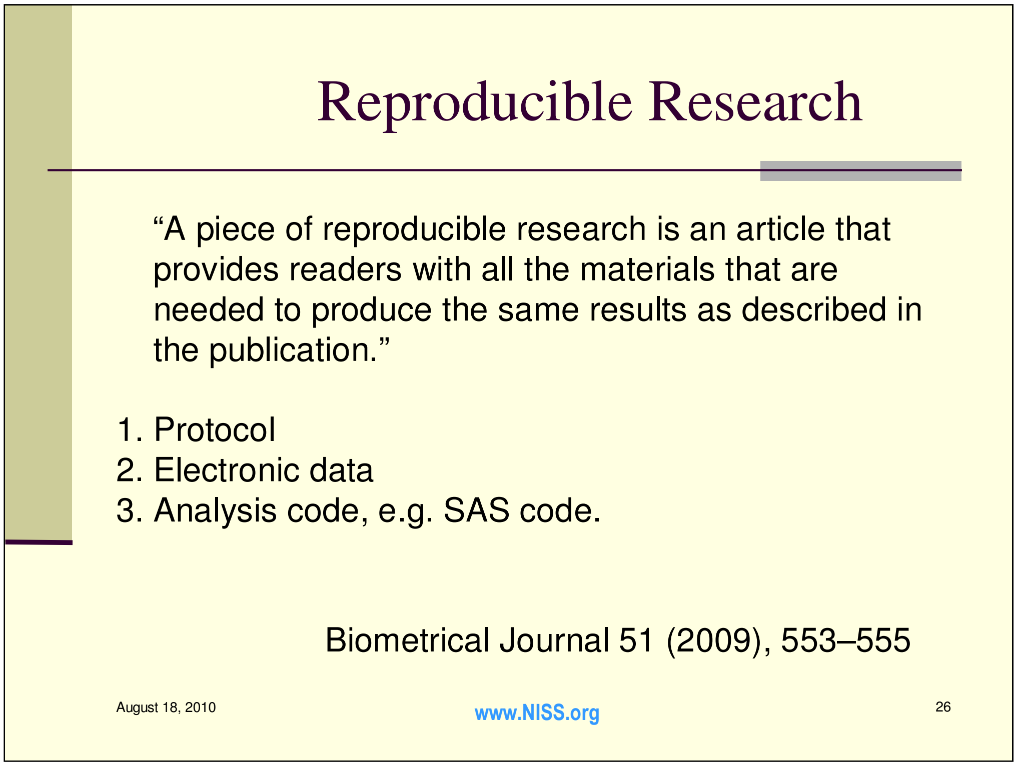
8. Family history
9. Personal history
10. Smoking

Etc.

August 18, 2010 www.NISS.org 25

There are a large number of potential confounders for a cardiovascular effect. Each of these confounders may be brought into and taken out of the model. There are ~1000 models in play.

Putting a few factors into a covariate adjustment model will look plausible. But how many models were tried? The reader will never know.



Reproducible Research

> “A piece of reproducible research is an article that provides readers with all the materials that are needed to produce the same results as described in the publication.”

1. Protocol
2. Electronic data
3. Analysis code, e.g. SAS code.

Biometrical Journal 51 (2009), 553–555

NB: this is a minimum start. It assumes that everything up to the production of the data set is sound. But see Feinstein, 1988. Starting with the data set, the protocol and the analysis code, the results in the paper are reproducible.