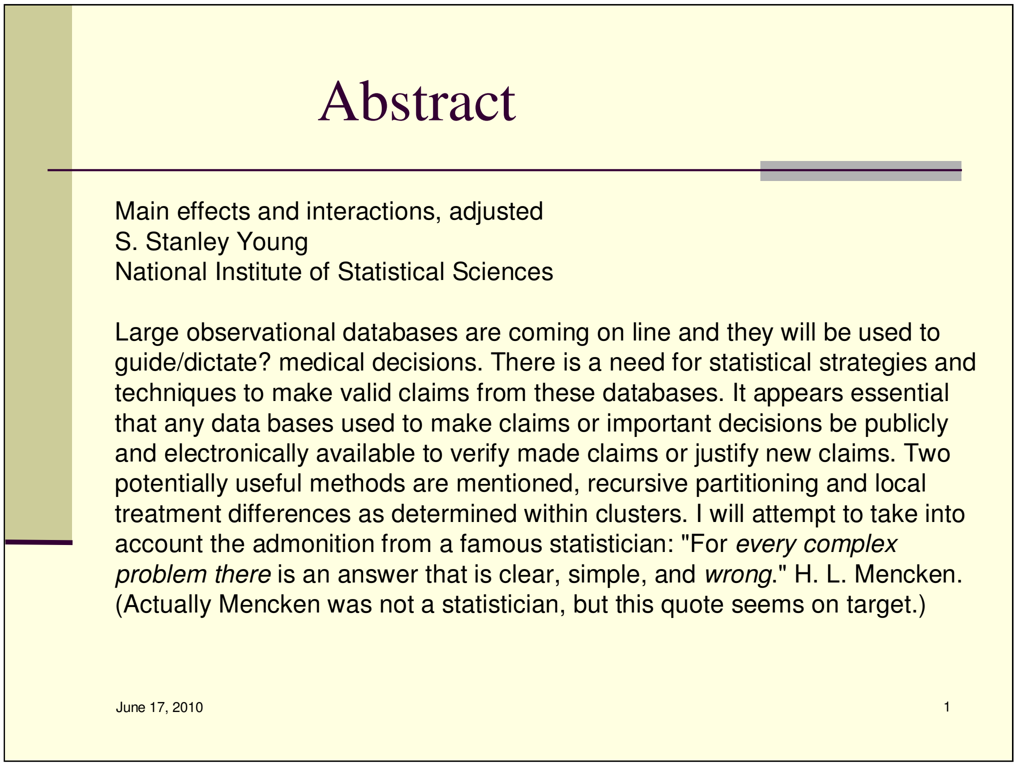


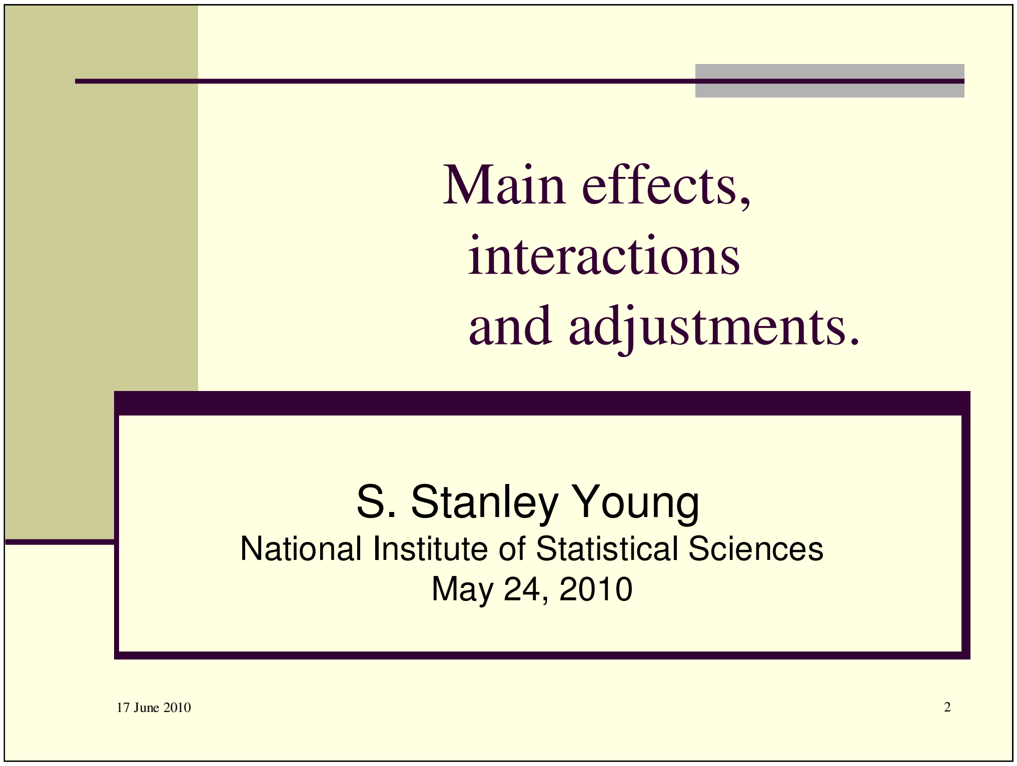
# Abstract

Main effects and interactions, adjusted

S. Stanley Young
National Institute of Statistical Sciences

Large observational databases are coming on line and they will be used to guide/dictate? medical decisions. There is a need for statistical strategies and techniques to make valid claims from these databases. It appears essential that any data bases used to make claims or important decisions be publicly and electronically available to verify made claims or justify new claims. Two potentially useful methods are mentioned, recursive partitioning and local treatment differences as determined within clusters. I will attempt to take into account the admonition from a famous statistician: "For *every complex problem there* is an answer that is clear, simple, and *wrong*." H. L. Mencken. (Actually Mencken was not a statistician, but this quote seems on target.)

Comparative Effectiveness Research, CER, has burst onto the scene with massive federal funding with the goal of attempting to figure out what medical procedures and drugs are most effective. There is interest in using observational studies to help in evaluations.



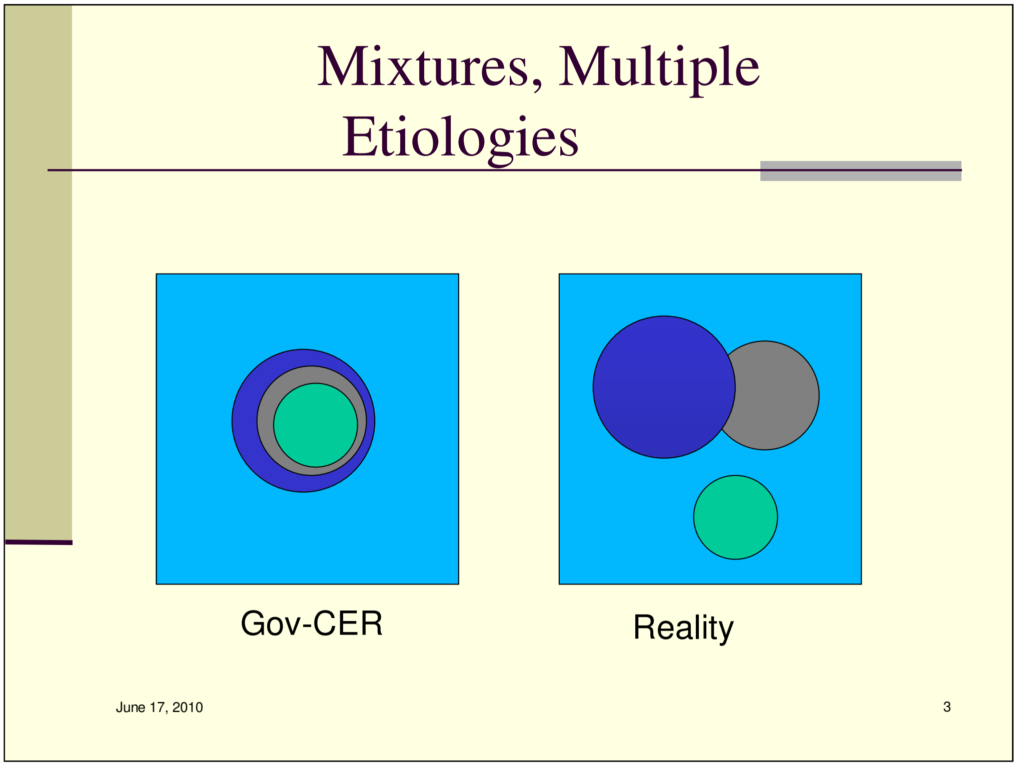
# Main effects, interactions and adjustments.

S. Stanley Young
National Institute of Statistical Sciences
May 24, 2010

17 June 2010

2

The simplest analysis is one looking at main effects, but for many medical situations this analysis is likely to be an oversimplification. Various factors will interact to make one treatment more effective in one situation and another treatment more effective in a different situation. Also, groups of patients will have different characteristics and there may be a need to statistically adjust the analysis to take these differences into account. Evaluation of large, observational studies is expected to be complex.



# Mixtures, Multiple Etiologies

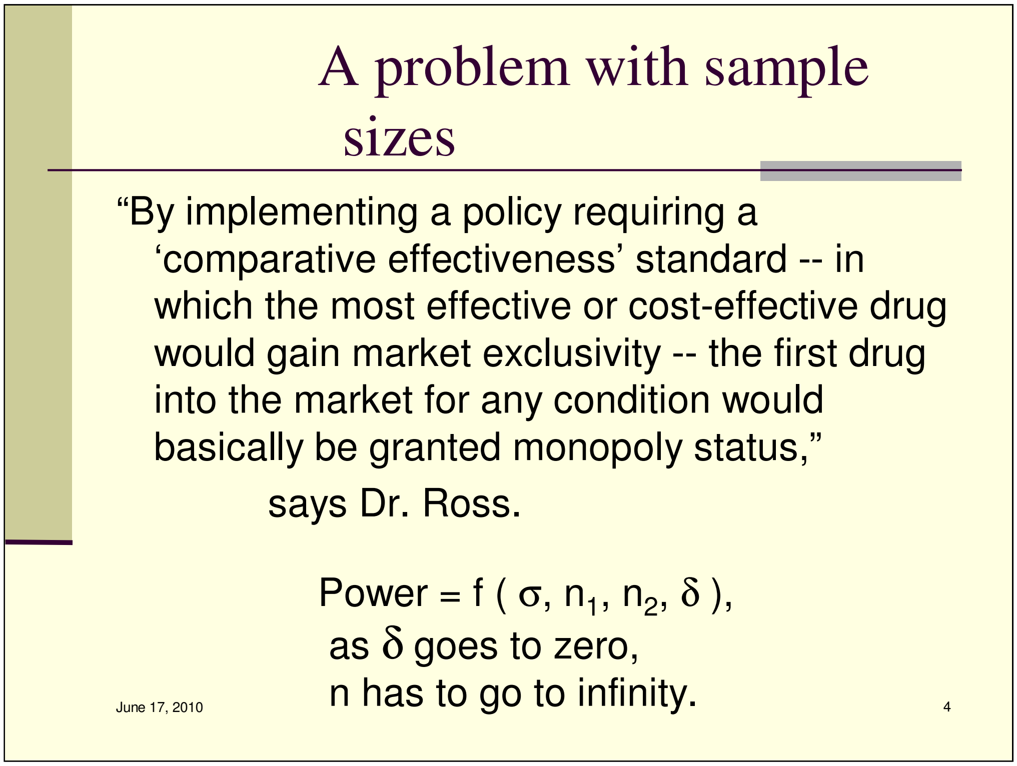
Gov-CER

Reality

June 17, 2010

3

In either situation, dark blue is the most effective treatment. Reality is likely much more complex. If only dark blue is funded under the 2nd scenario, then many people that could be effectively treated will not be treated with drugs that work for them.



# A problem with sample sizes

"By implementing a policy requiring a 'comparative effectiveness' standard -- in which the most effective or cost-effective drug would gain market exclusivity -- the first drug into the market for any condition would basically be granted monopoly status,"

says Dr. Ross.

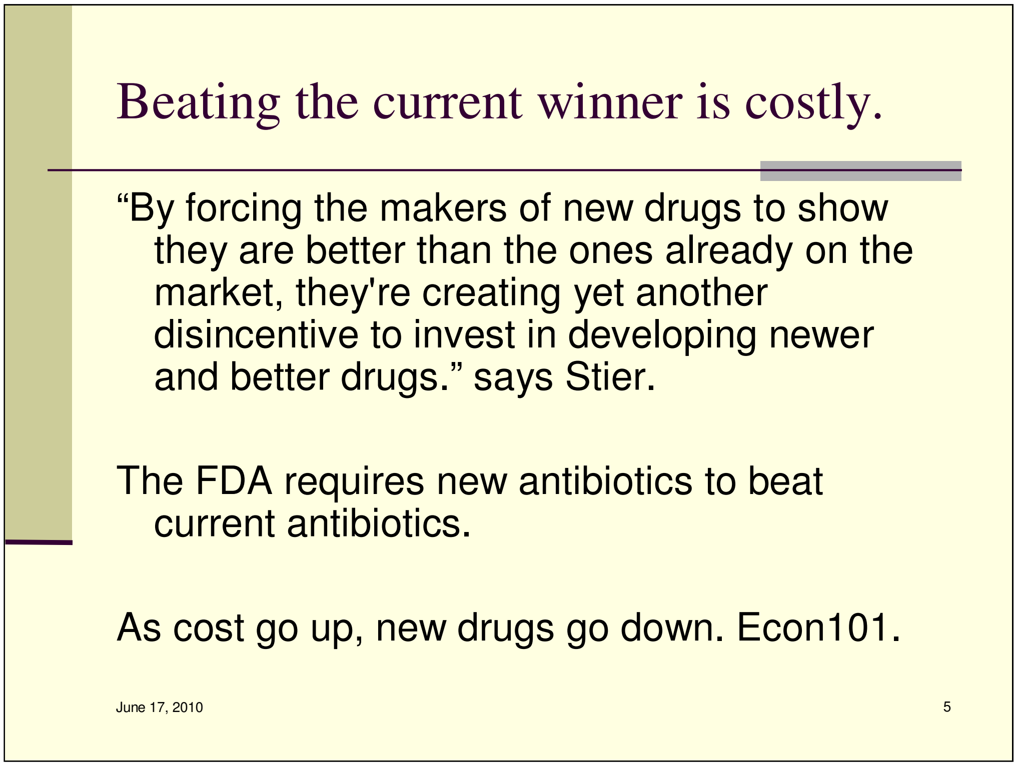
Power = f ( $\sigma$, $n_1$, $n_2$, $\delta$ ),
as $\delta$ goes to zero,
n has to go to infinity.

June 17, 2010

4

One unintended consequence of requiring that any new drug would have to beat out the standard drug in a head to head trial would be the likely decrease in bringing new drugs to market. The expected difference in effectiveness of a drug to placebo would be much larger that the difference to an existing drug so the sample size would have to be much larger. The sample size would likely be so much larger that a rational person would not spend the funds to conduct the trial.

The first person to market for an indication would likely be the last person to market.



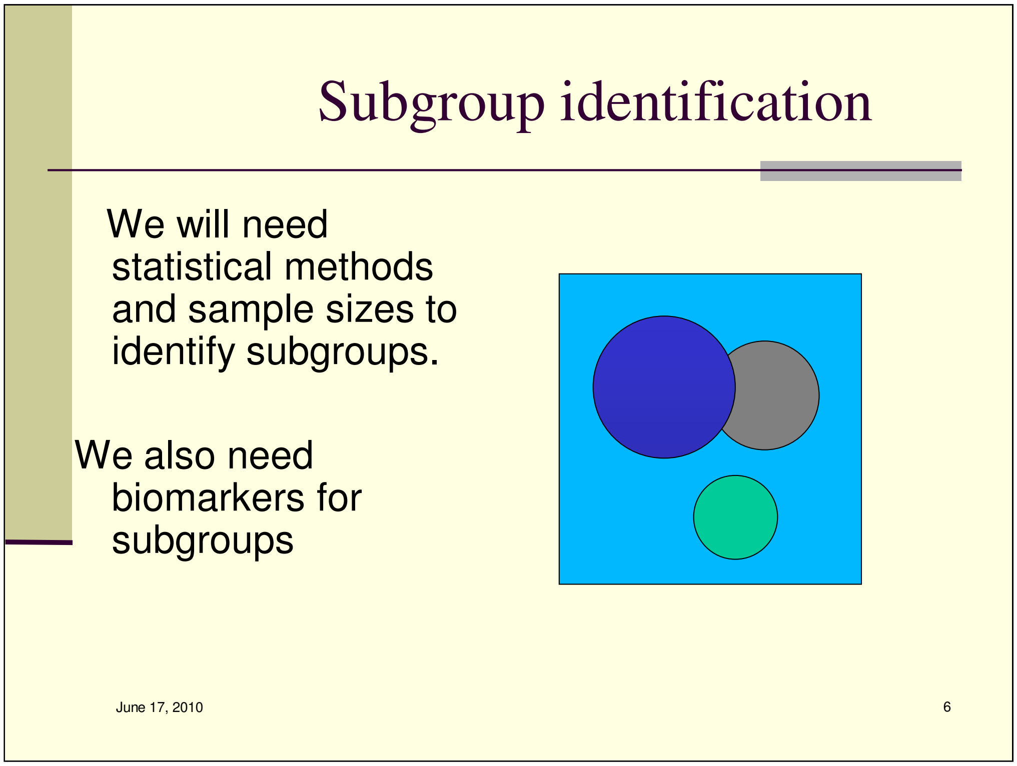
## Beating the current winner is costly.

"By forcing the makers of new drugs to show they are better than the ones already on the market, they're creating yet another disincentive to invest in developing newer and better drugs." says Stier.

The FDA requires new antibiotics to beat current antibiotics.

As cost go up, new drugs go down. Econ101.

Clearly as the cost of bringing a drug to market go up, there will be fewer drugs brought to market.



# Subgroup identification

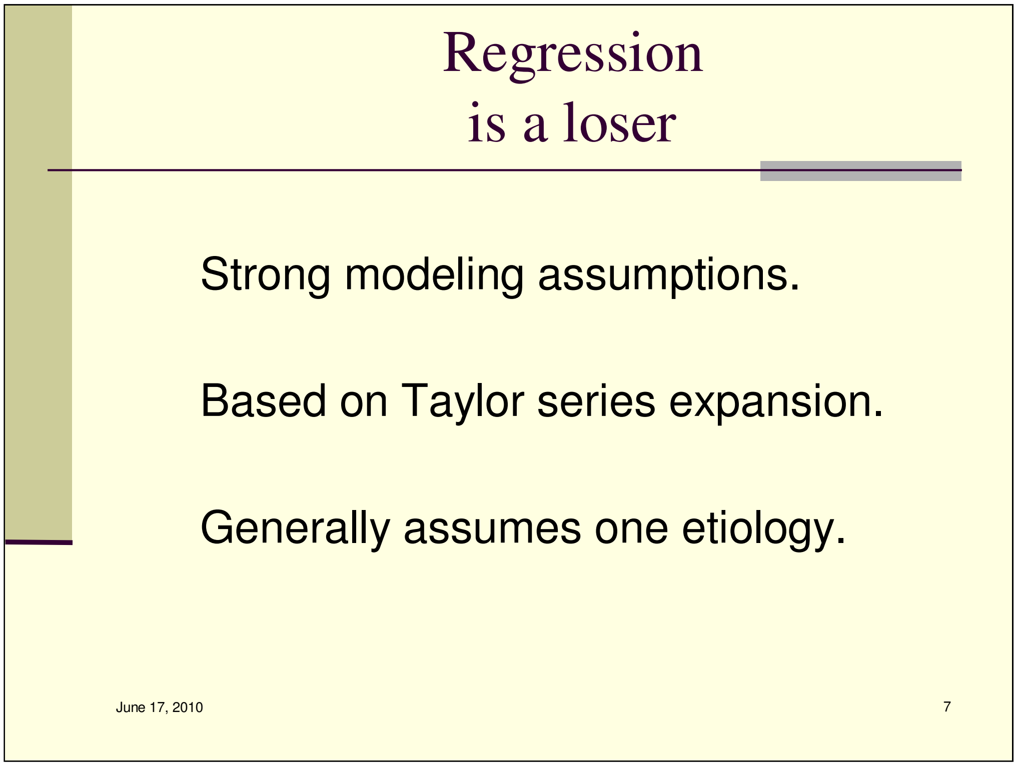
We will need statistical methods and sample sizes to identify subgroups.

We also need biomarkers for subgroups

June 17, 2010

6

Finding subgroups of patients where a drug is effective is or will become important. It is likely that many named diseases are really made up of multiple diseases where patient characteristics or etiologies are different. We will need biomarkers to identify the subgroups of patients where different drugs are effective.



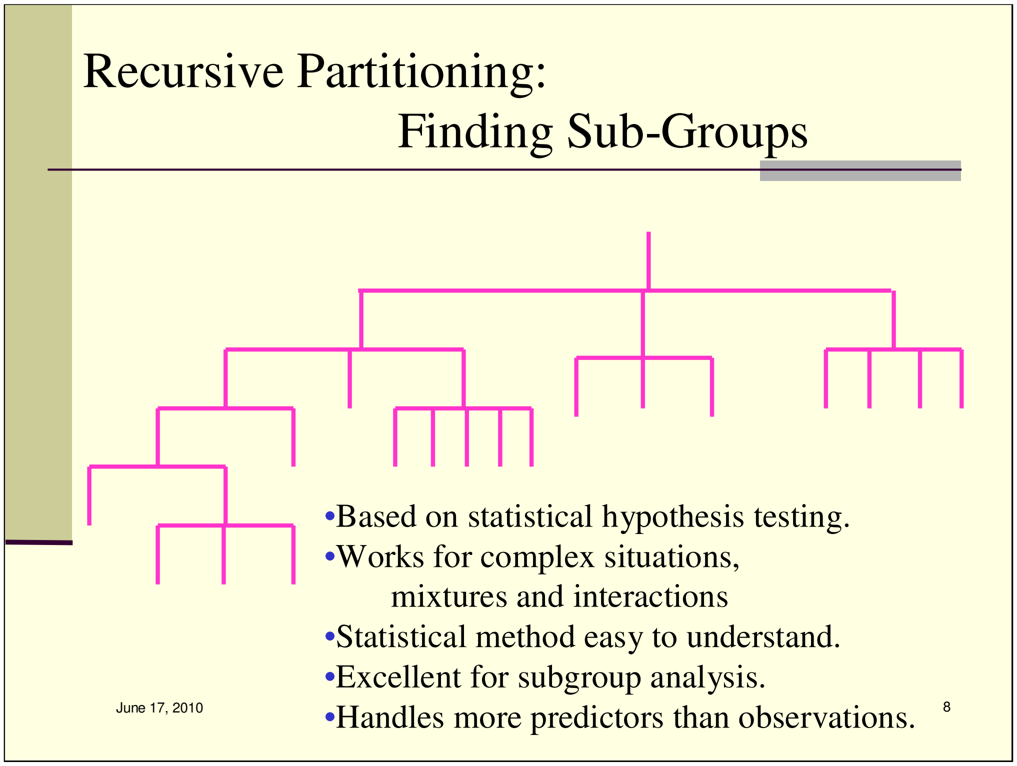
# Regression is a loser

Strong modeling assumptions.

Based on Taylor series expansion.

Generally assumes one etiology.

Regression is a loser from many points of view. The biggest problem is that a single named disease may have multiple etiologies. Predictors that are important for one etiology and unlikely to be important for another etiology. Regression coefficients will be an average over the etiologies and hence are likely to be very misleading.



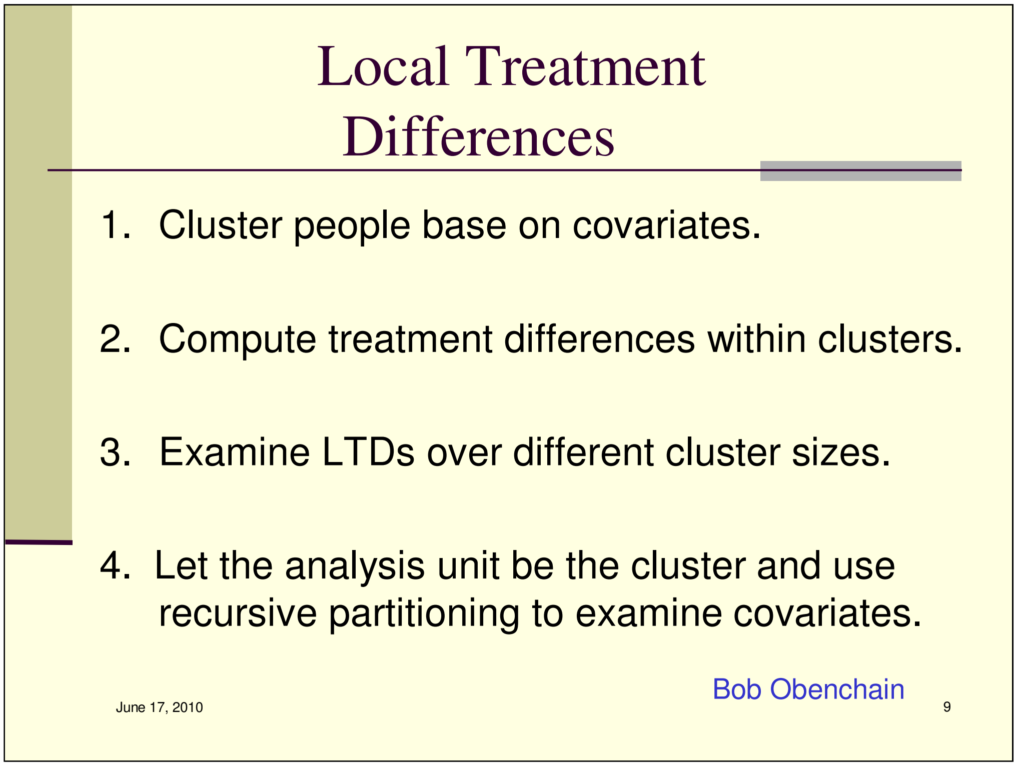
## Recursive Partitioning: Finding Sub-Groups

- Based on statistical hypothesis testing.
- Works for complex situations, mixtures and interactions
- Statistical method easy to understand.
- Excellent for subgroup analysis.
- Handles more predictors than observations.

June 17, 2010

8

Recursive partitioning could be a big winner as it can handle mixture situations. In my opinion, the algorithms of Doug Hawkins are the best. Splitting is based on multiplicity adjusted multiple testing p-values. See Hawkins DM. Recursive partitioning. *Computational Statistics*. 2009;1:290-295 and the references therein.



## Local Treatment Differences

1. Cluster people base on covariates.
2. Compute treatment differences within clusters.
3. Examine LTDs over different cluster sizes.
4. Let the analysis unit be the cluster and use recursive partitioning to examine covariates.

Bob Obenchain

June 17, 2010

9

Blocking is widely used in the statistics world. Here that means grouping people together that are similar. Within a cluster treatment differences can be computed and those differences are "adjusted" for the variables used in the clustering. The distribution of "local treatment differences" can be examined as the number of clusters are varied. Obviously, with only one cluster you are looking at the "main effects" and with everyone in their own cluster, no differences can be computed. There will be uninformative clusters, clusters that have only one treatment. There is usually a number of clusters where the LTD distribution stabalizes.

Finally, one can use modeling to examine the influence of covariates on the LTDs.

See

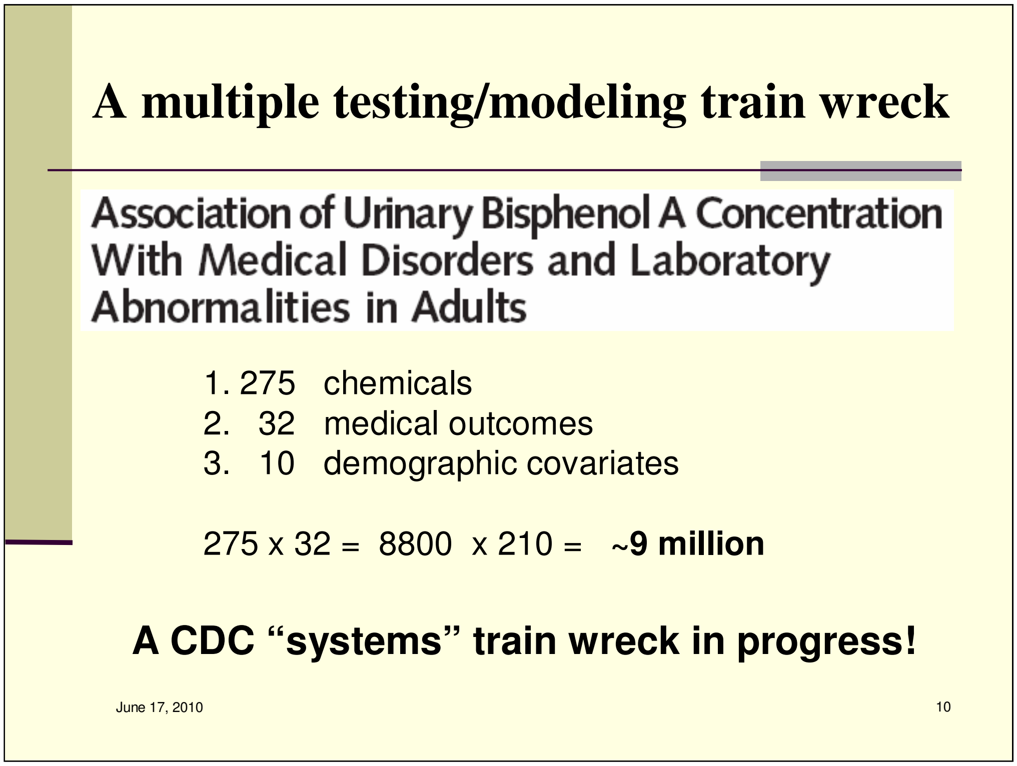
**Robert L. (Bob) Obenchain, PhD, FASA**

**Risk Benefit Statistics LLC.**

13212 Griffin Run, Carmel, IN 46033-8835

(317) 580-0144; softrx@iquest.net

Obenchain RL. Identifying Meaningful Patient Subgroups via Clustering – Sensitivity Graphics. **2006 JSM Proceedings on CD-ROM.** (6 pages.) Alexandria, VA: American Statistical Association. 2007.



# A multiple testing/modeling train wreck

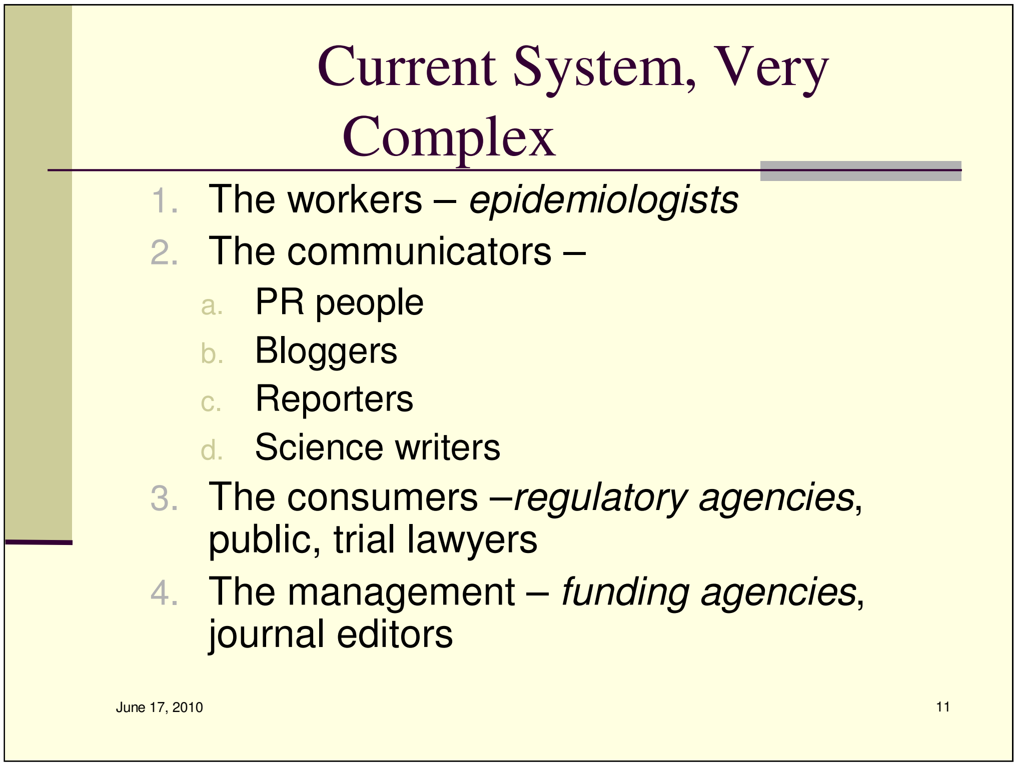
Association of Urinary Bisphenol A Concentration With Medical Disorders and Laboratory Abnormalities in Adults

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

275 x 32 = 8800 x 210 = ~**9 million**

**A CDC “systems” train wreck in progress!**

Two problems with the analysis of observational studies is not taking account of multiple testing and multiple model building. Basically, extensive searching through the data set can find things that look unusual, but are false positives – the finding will not replicate. This paper appeared in JAMA and the claims are most likely false. Sorting through about 9 million items, you will undoubtedly find things that look unusual. Remember, it is up to the person making a claim to provide supporting evidence.



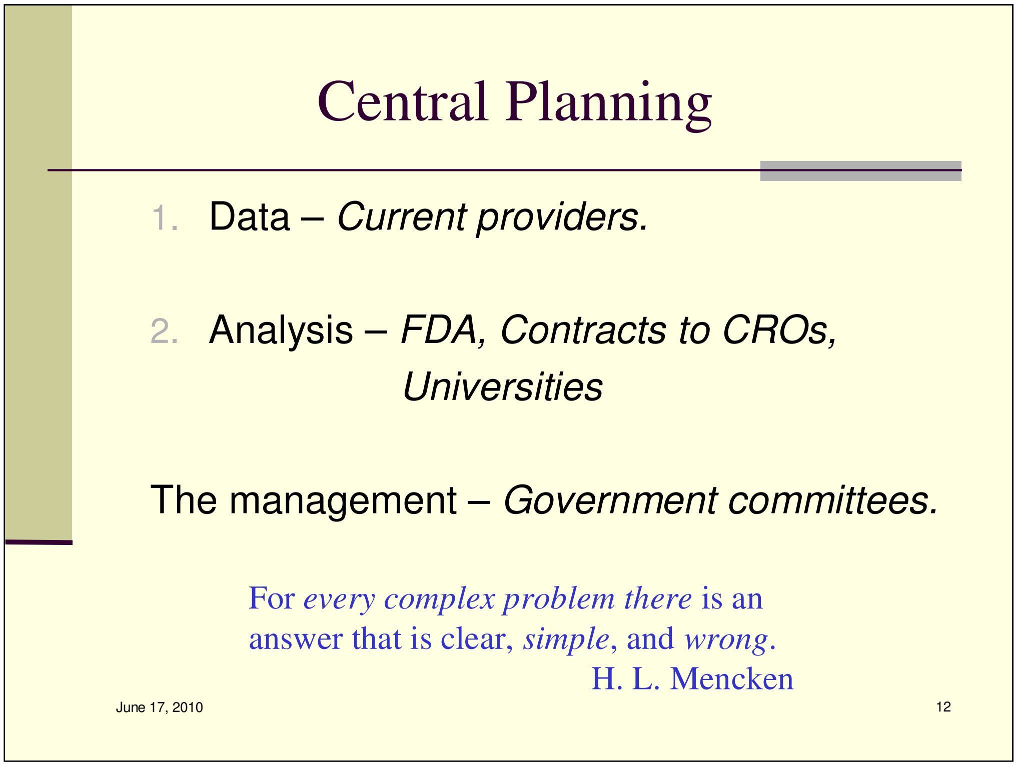
## Current System, Very Complex

1. The workers – *epidemiologists*
2. The communicators –
   a. PR people
   b. Bloggers
   c. Reporters
   d. Science writers
3. The consumers –*regulatory agencies*, public, trial lawyers
4. The management – *funding agencies*, journal editors

June 17, 2010 11

It is normal to frame this problem with two players, the epidemiologists and the consumers. In the case of CER, epidemiologists analyze large observational and clinical trial data sets, and make claims. Government committees set funding policy for government and insurance payment plans. We think the problem is more complicated. The papers contain claims and those claims are re-packaged and communicated to the consumers. There is quite typically a “press release” for a paper describing results from an observational study. Usually, the press release omits important caveats and warnings mentioned in the paper. Warnings in the paper are usually “in code” and toward the end of the paper. The internet is now the first and primary medium of communications. Reporters will have two weeks while the paper is embargoed to get their stories together. All this, paper, press release, web reports, stories hits the consumer at one time – the official release date of the paper. It can be difficult for the consumers to react. The message can be loud and multi-media. A letter to the editor can take months to appear. By that time, the reporters have largely moved on. Blogging is possible, but can be diffuse. A myth can be made before effective counters appear.

The managers are largely out of the loop. The system is in place and functions largely on its own. IF the system is out of control, they have to change the system. Workers and communications people are executing the existing system. Consumers are largely “on their own” in responding to claims from observational studies.



## Central Planning

1. Data – *Current providers.*
2. Analysis – *FDA, Contracts to CROs, Universities*

The management – *Government committees.*

> For *every complex problem there* is an answer that is clear, *simple*, and *wrong*.
> H. L. Mencken

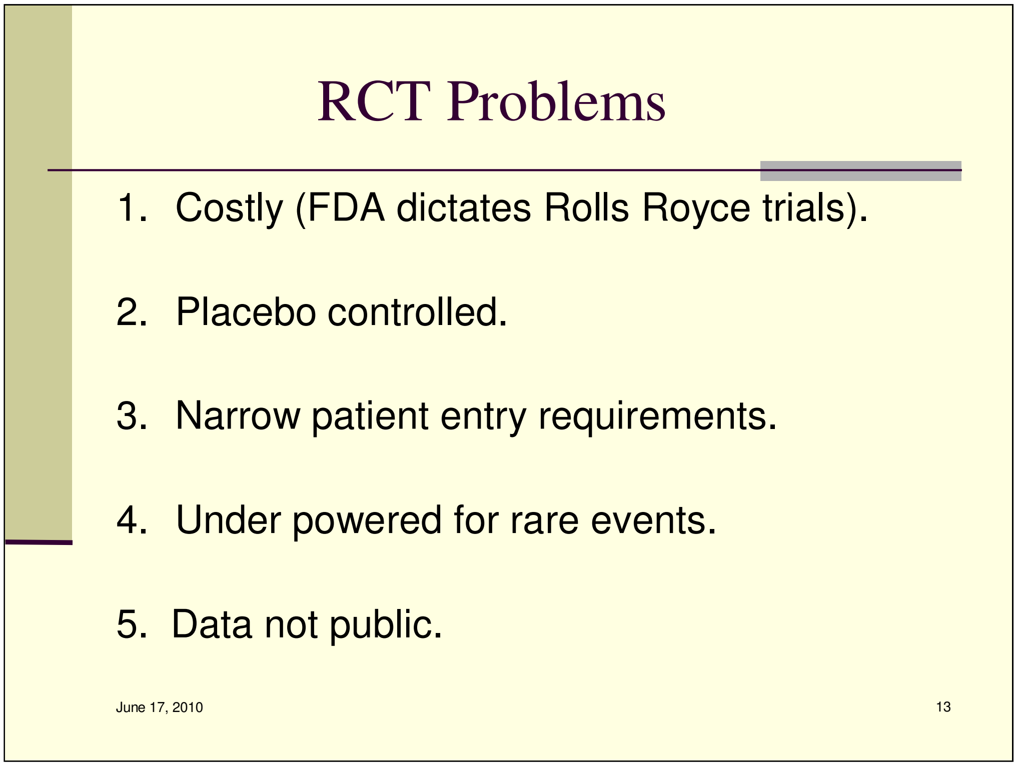
June 17, 2010 12

Currently, the data is owned by health care providers. They view the data as a profit center and as a source to figure out how to control their costs.

Congress has charged the FDA to construct a large observational data base to be able to detect rare side effects and also possibly use for CER.

Access to data by all affected parties appears to be critical to have some level of oversight.

Heath care in the US is very complex. It is not clear that central committees are capable of managing such a system effectively.



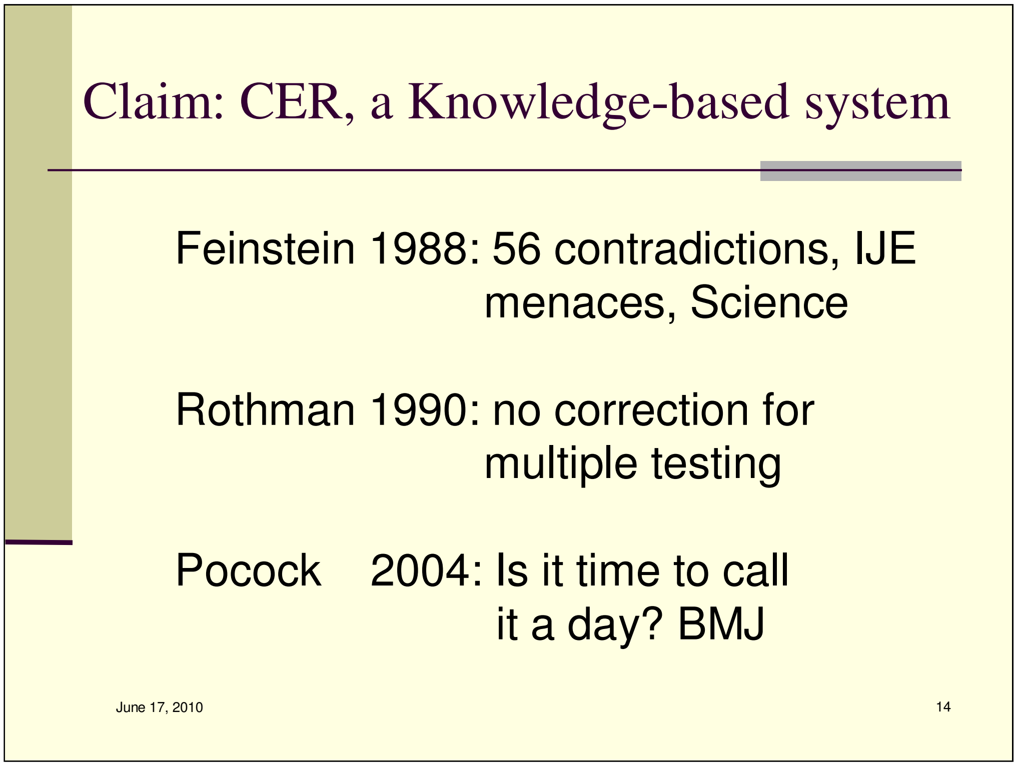
## RCT Problems

1. Costly (FDA dictates Rolls Royce trials).
2. Placebo controlled.
3. Narrow patient entry requirements.
4. Under powered for rare events.
5. Data not public.

June 17, 2010 13

Randomized Clinical Trials are not without their problems, listed here. But the do move us forward with relatively sure knowledge. Claims from RCT replicate ~80% of the time. Given power considerations, that means they are correct a high percentage of the time.

They are likely much more costly than they need to be when the process is dictated by the FDA.



# Claim: CER, a Knowledge-based system

Feinstein 1988: 56 contradictions, IJE menaces, Science

Rothman 1990: no correction for multiple testing

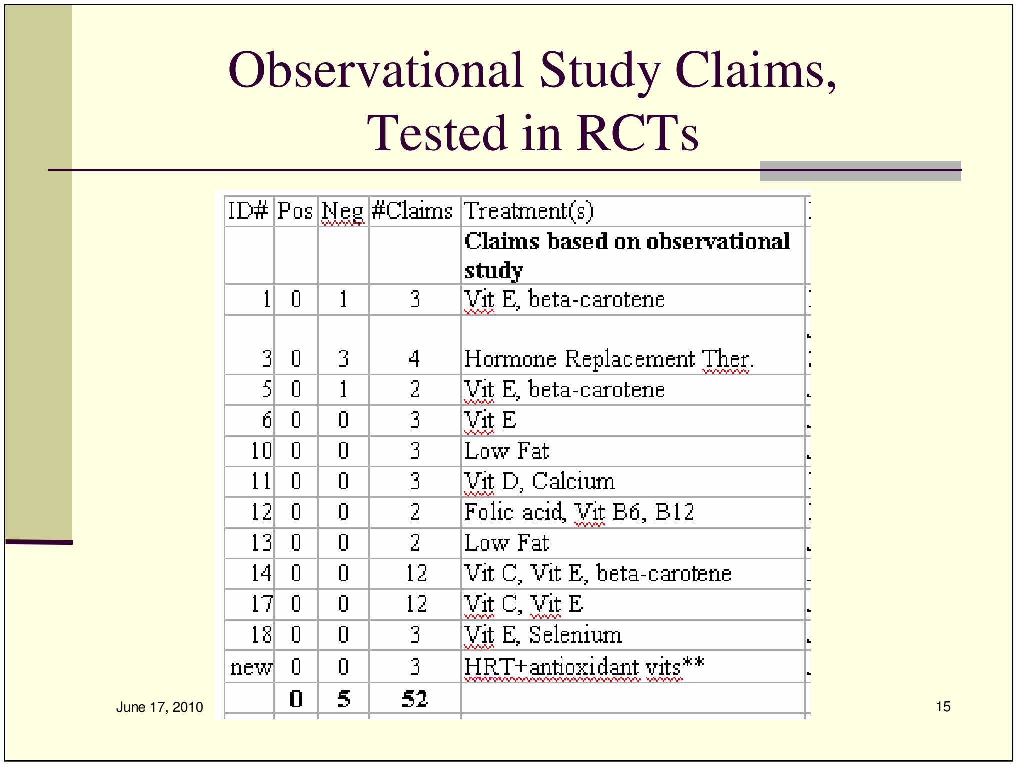
Pocock 2004: Is it time to call it a day? BMJ

June 17, 2010

14

Current knowledge comes from clinical trials, where there are two problems. First, most trials are placebo controlled so there is no good way to compare one treatment with another. Second, patient entry requirements can be very restricted so it is not known how well successful treatments will generalize.

Current knowledge comes from observational studies where the there is essentially no oversight and claims fail to replicate over 90% of the time.



# Observational Study Claims, Tested in RCTs

| 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** | |

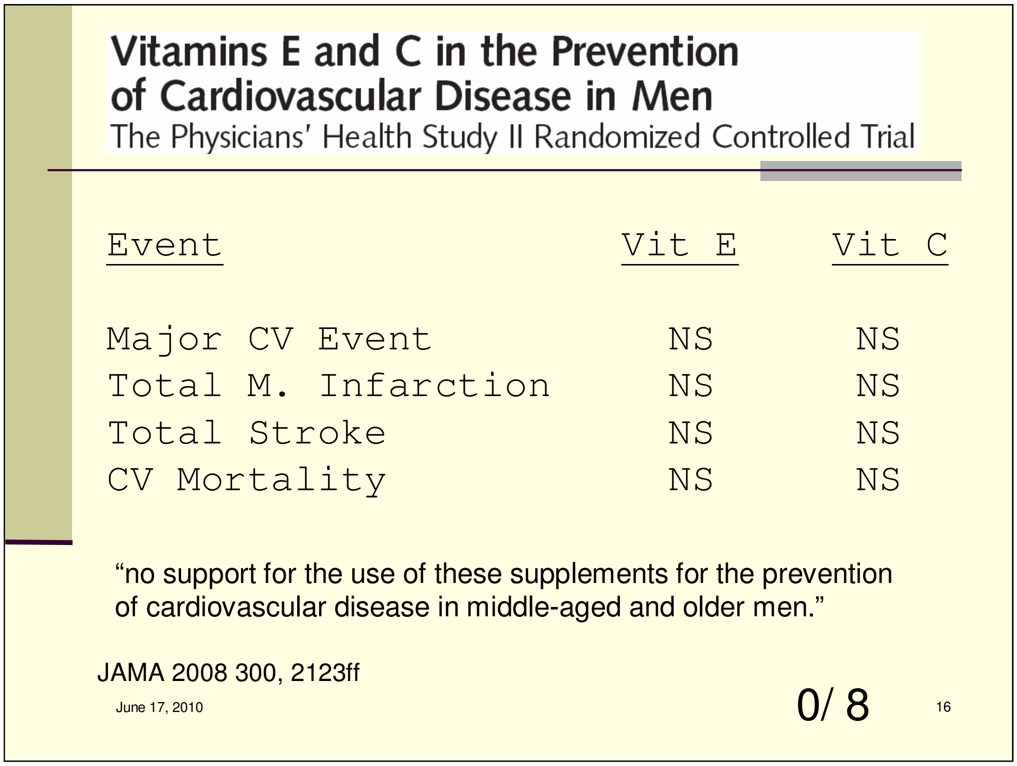
June 17, 2010

15

The NIH has funded many studies to test in RCTs the claims coming from observational studies. Of 52 claims tested, not one has been confirmed. 5 claims were statistically significant in the opposite direction from the observational study claim.

Observational studies are out of control. There is no oversight of observational studies.

Yet, CER will depend heavily on observational studies!



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

The Physicians' Health Study II Randomized Controlled Trial

| Event | Vit E | Vit C |
|---|---|---|
| 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

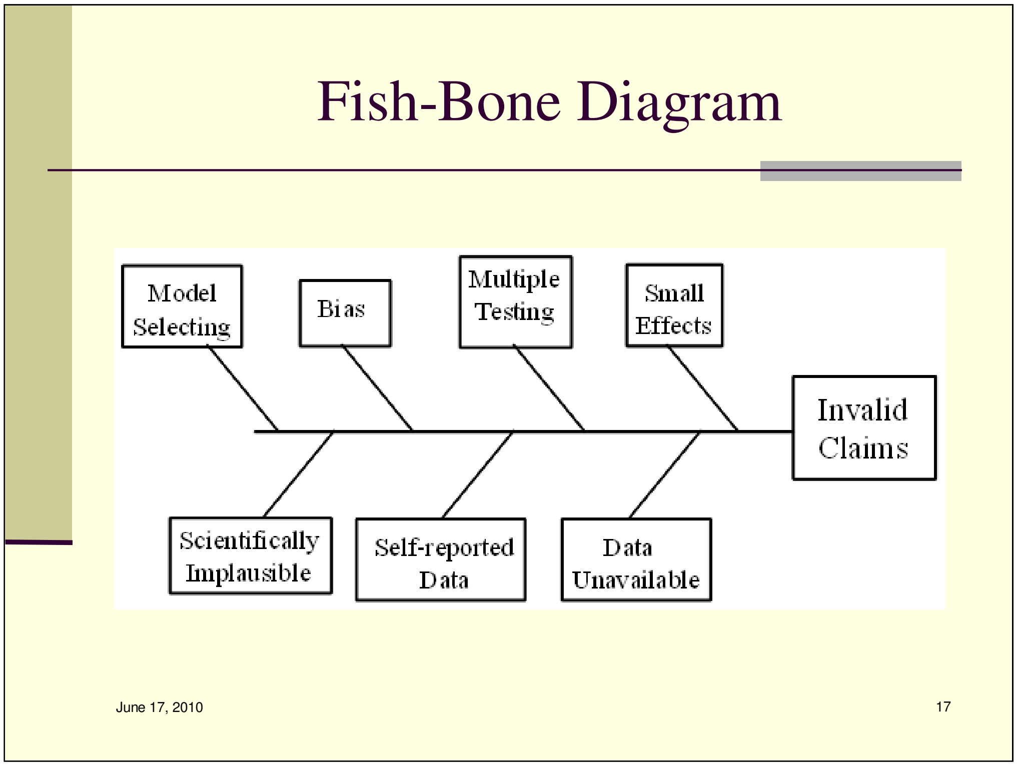
June 17, 2010

0/ 8

16

Vitamins E and C have been repeatedly given rise to "protective" claims in observational studies. Elaborate "anti-oxidant" rationalizations have been put forth.

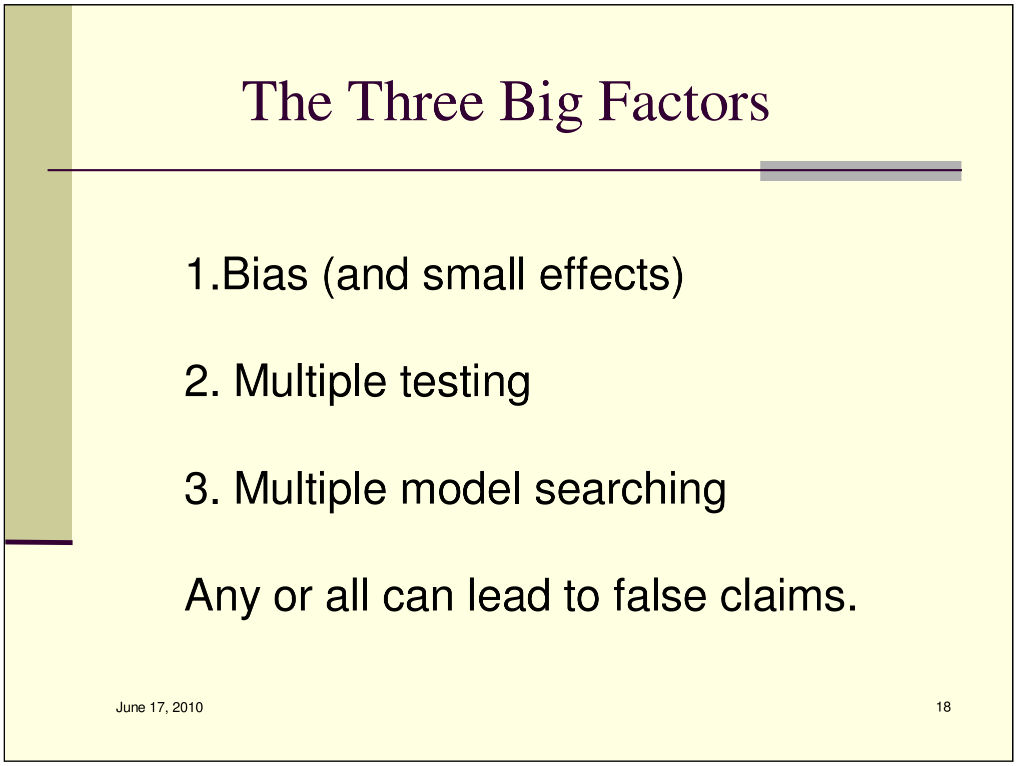
In a RCT Vits E and C are 0 for 8 on replicating claims! None of the claims replicated.



# Fish-Bone Diagram

June 17, 2010 17

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.



## The Three Big Factors

1.Bias (and small effects)

2. Multiple testing

3. Multiple model searching

Any or all can lead to false claims.

June 17, 2010 18

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

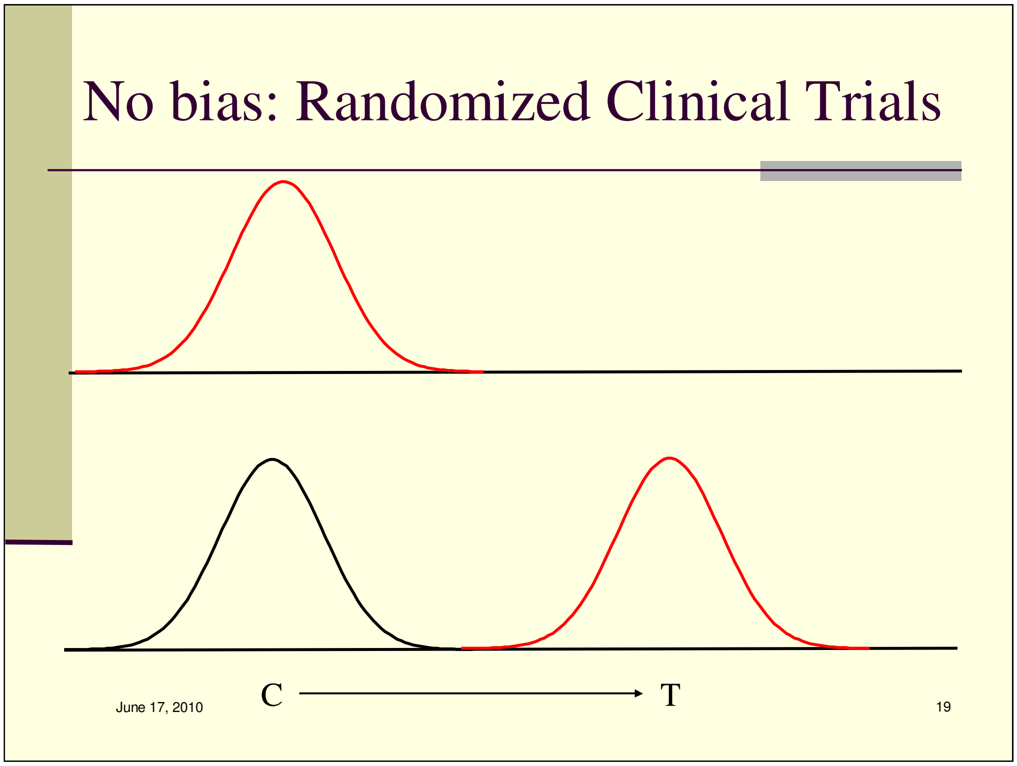
The word bias covers a lot of sins. Unmeasured confounders. Measure, but unused confounders. Modeling bias – run hundreds of models and select the one you like.

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 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.



## No bias: Randomized Clinical Trials

June 17, 2010

19

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

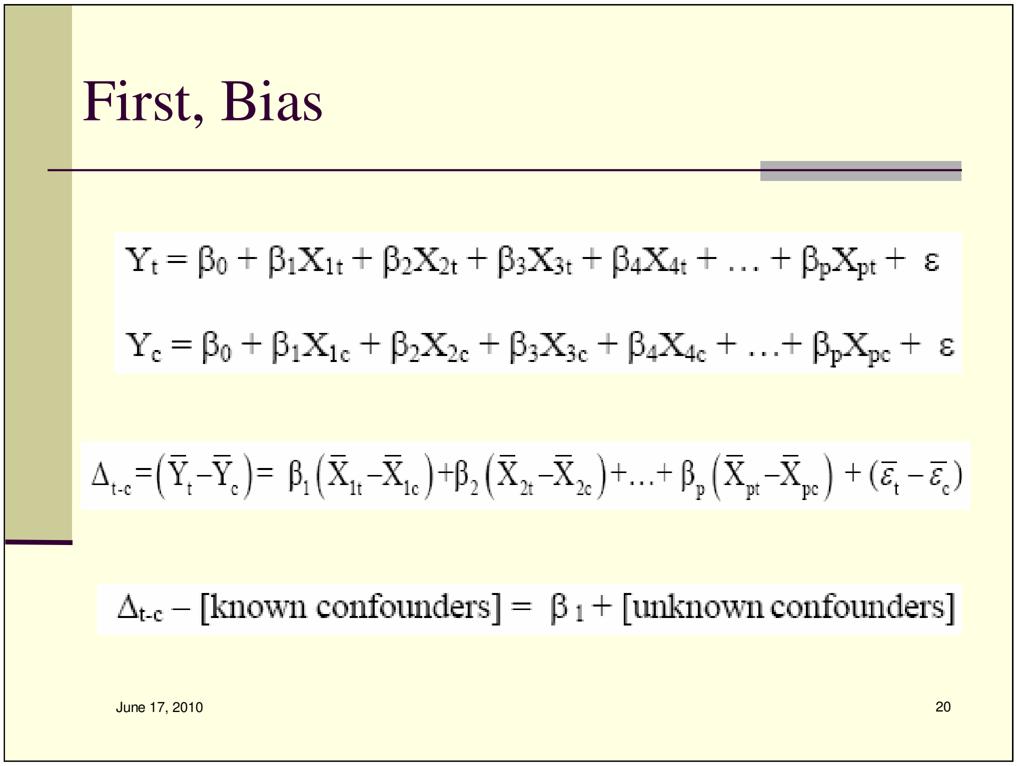
If treatment has an effect it will move the distribution of the treated patients away from the control patients. If the effect is large enough and if the sample size is large enough, the treatment effect will be detected.

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.



# First, Bias

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

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

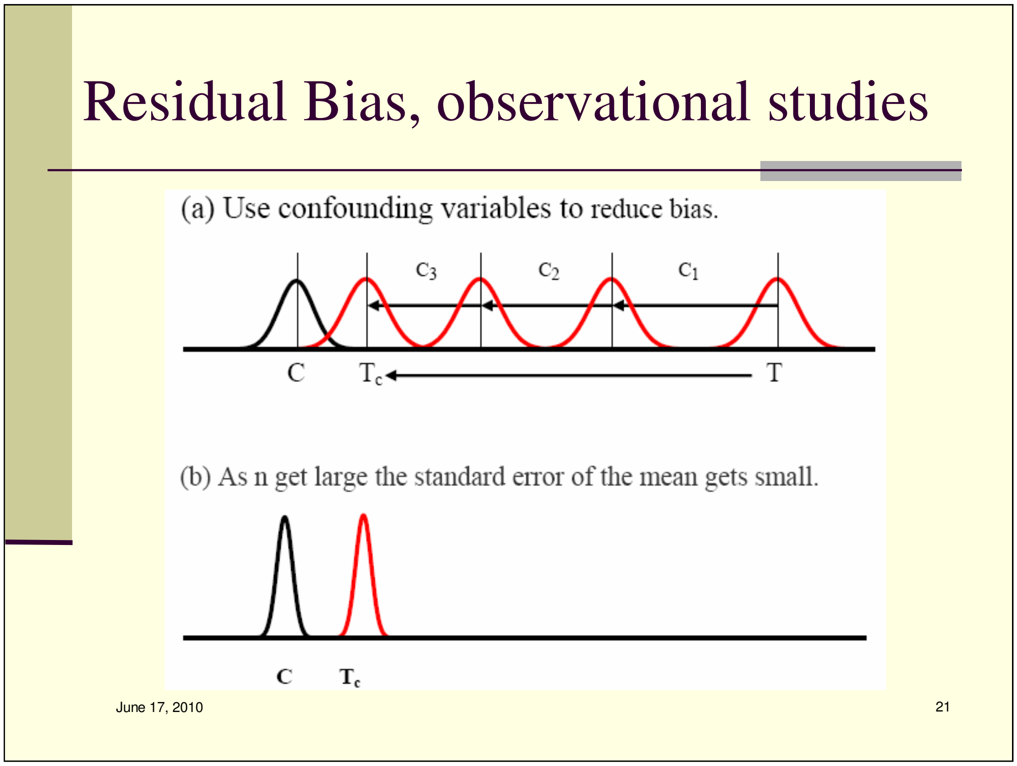
$$\Delta_{t\text{-}c} = \left(\bar{Y}_t - \bar{Y}_c\right) = \beta_1\left(\bar{X}_{1t} - \bar{X}_{1c}\right) + \beta_2\left(\bar{X}_{2t} - \bar{X}_{2c}\right) + \ldots + \beta_p\left(\bar{X}_{pt} - \bar{X}_{pc}\right) + (\bar{\varepsilon}_t - \bar{\varepsilon}_c)$$

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

June 17, 2010

20

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 and take the value 0. 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. We think we are looking at Beta1 when we are really looking at Beta1 along with all unknown confounders.



# Residual Bias, observational studies

(a) Use confounding variables to reduce bias.

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

June 17, 2010

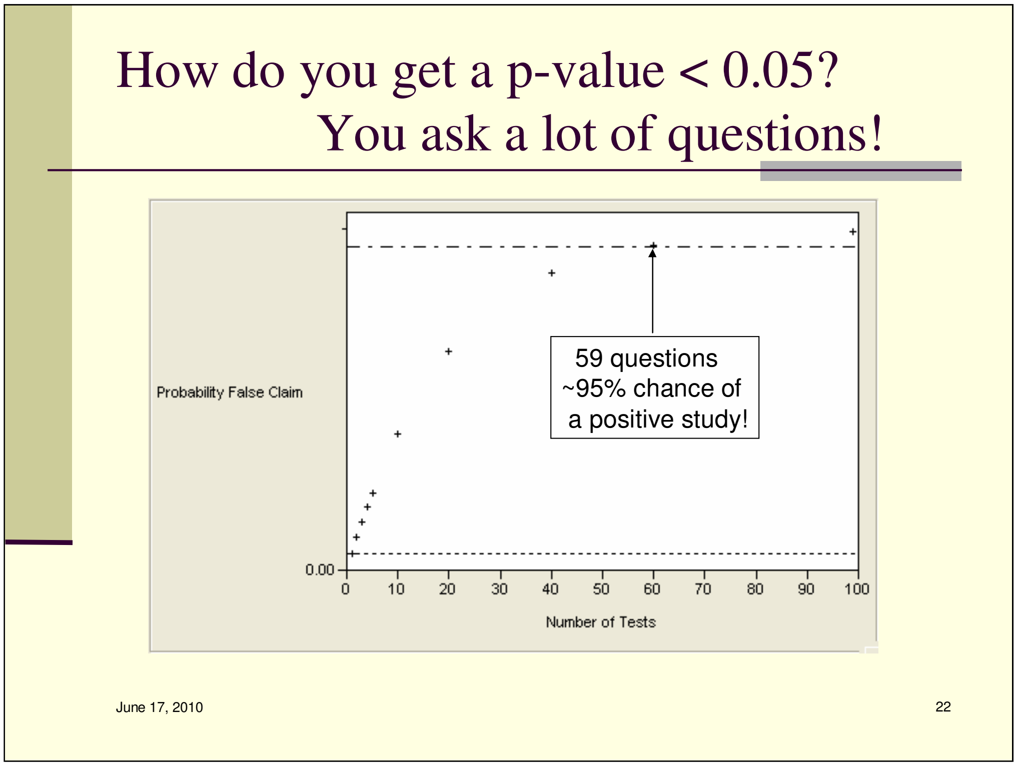
21

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 5-10 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.



# How do you get a p-value < 0.05? You ask a lot of questions!

59 questions
~95% chance of
a positive study!

June 17, 2010

22

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

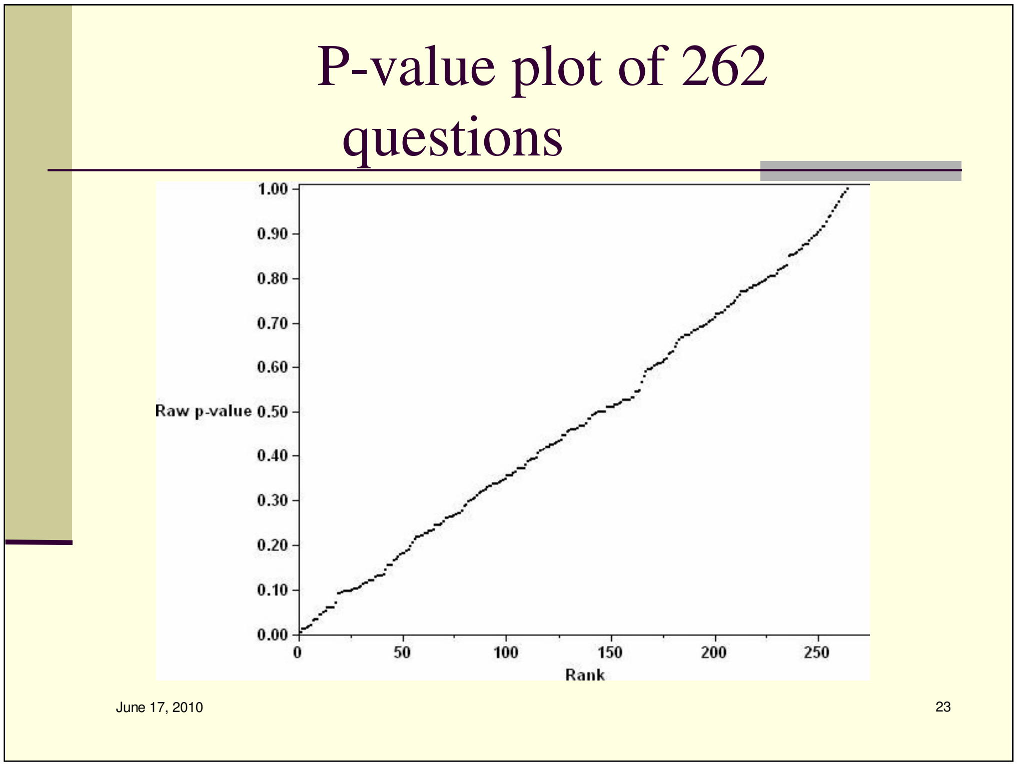
If 59 independent questions are asked in an experiment there is a 95% probability of at least one “statistically significant” result.

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.

It is typical in observational studies to ask a lot of questions and not clearly state how many questions are at issue. It is not easy for the reader to count the questions.

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



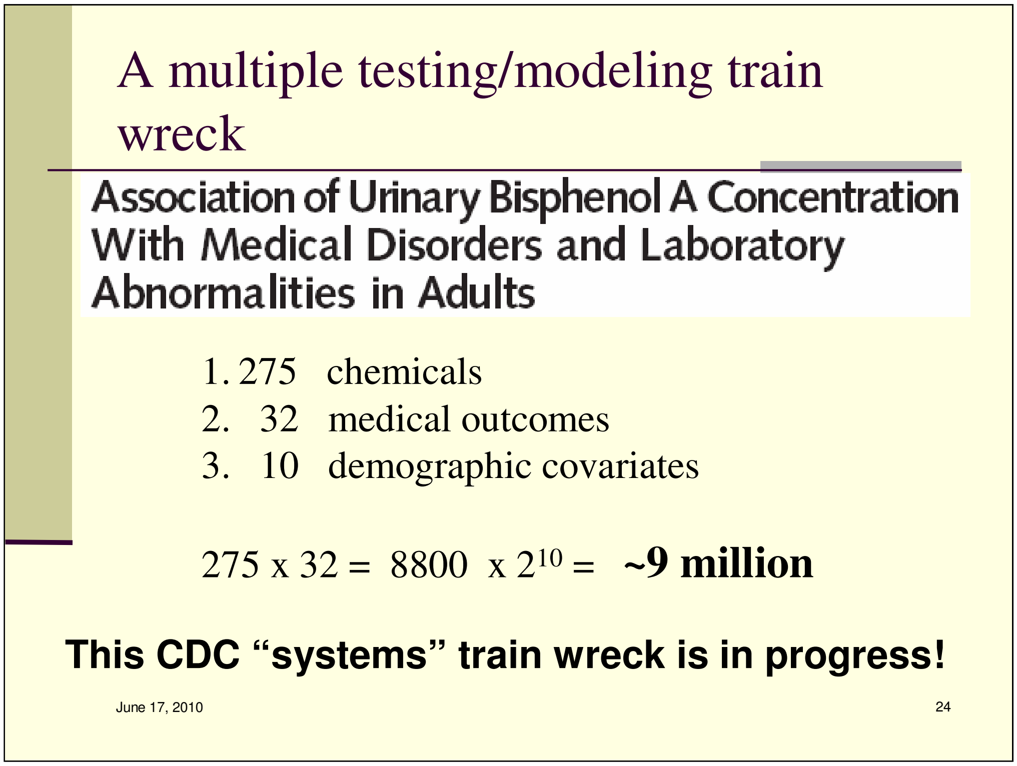
## P-value plot of 262 questions

June 17, 2010

23

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.



## A multiple testing/modeling train wreck

**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**

**This CDC "systems" train wreck is in progress!**

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.

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.

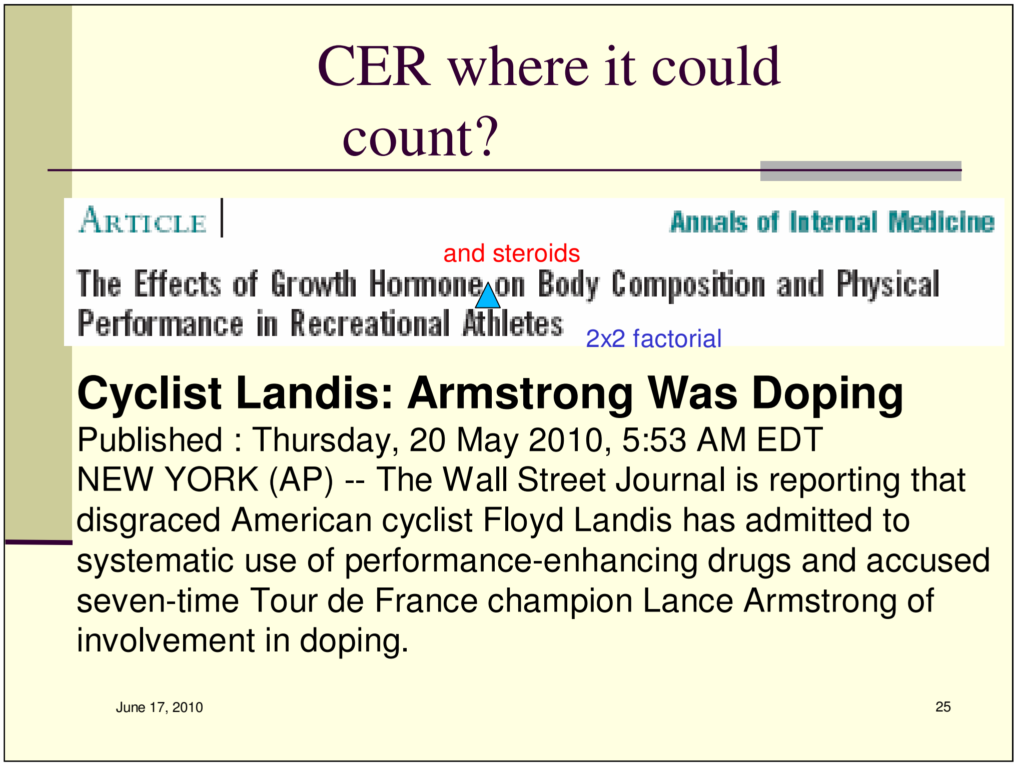
The CDC NHANES (2003-2004) measured 275 environmental chemicals and a wide range of health outcomes. Although the study by Lang et al focused on 1 chemical and 16 health outcomes (8 patient-reported medical outcomes and 8 clinical chemistry measurements), counting to determine how many questions were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



## CER where it could count?

ARTICLE | Annals of Internal Medicine

The Effects of Growth Hormone and steroids on Body Composition and Physical Performance in Recreational Athletes 2x2 factorial

**Cyclist Landis: Armstrong Was Doping**

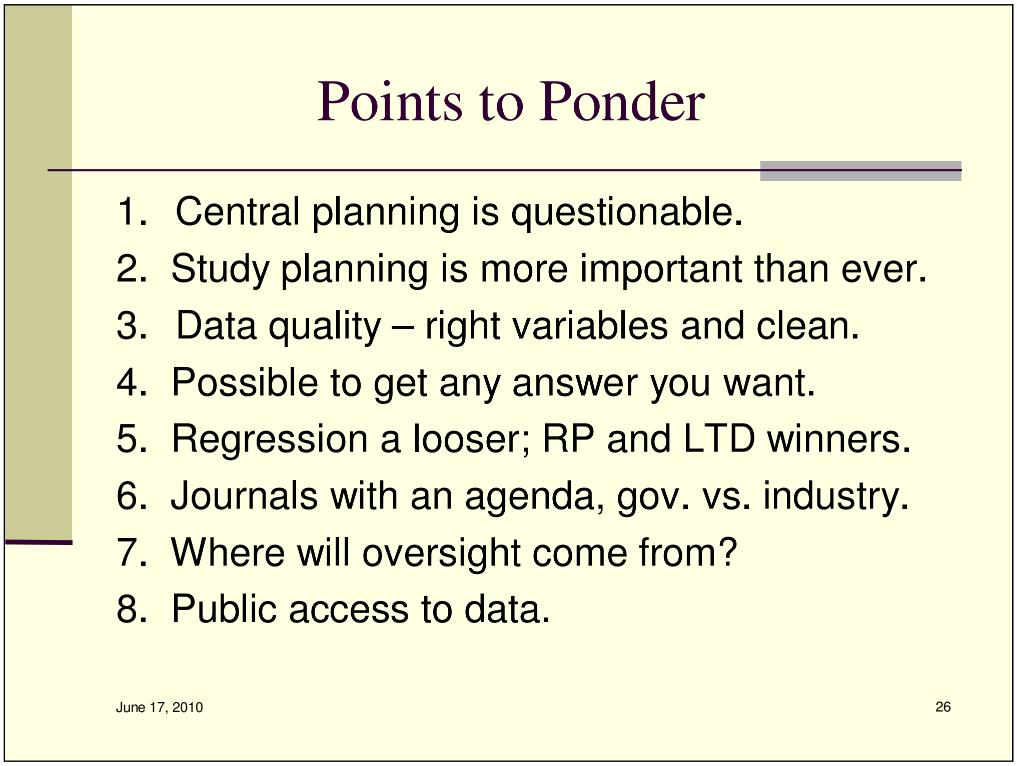
Published : Thursday, 20 May 2010, 5:53 AM EDT

NEW YORK (AP) -- The Wall Street Journal is reporting that disgraced American cyclist Floyd Landis has admitted to systematic use of performance-enhancing drugs and accused seven-time Tour de France champion Lance Armstrong of involvement in doping.

June 17, 2010 25

There are non-approved uses of drugs that rather obviously work. Athletes have been testing and using drugs for decades, often with state sponsorship.

On the other hand, natural products appear to get a free ride in the US.



# Points to Ponder

1. Central planning is questionable.
2. Study planning is more important than ever.
3. Data quality – right variables and clean.
4. Possible to get any answer you want.
5. Regression a looser; RP and LTD winners.
6. Journals with an agenda, gov. vs. industry.
7. Where will oversight come from?
8. Public access to data.

June 17, 2010

26

WW II was fought in large part to protect the individual from central planners. The implosion of Russia seems to have settled the question of economic efficiency.

Journals often have an agenda.

Big government and big science warrant some concern.

Often there is essentially no oversight of claims in a paper. Public access to data offers some oversight.

# Contact Information

Stan Young
young@niss.org
www.niss.org
Research/analysis

www.Omicsoft.com
stan.young@omicsoft.com
Custom programming - LTD

June 17, 2010 27

Dealing with large observational data set arguably requires specialized software as well as new thinking. In particular, systems for computing and visualizing local treatment differences is a need.