SCS 2017: Longitudinal and Nested Data

Day 1 – Part 1

The Fundamental 2 x 2 Table of Statistics

– or when Statistics seem to lie they are usually just answering a different question

Georges Monette, March 2017

There are three kinds of lies:

lies, damned lies and statistics

– Benjamin Disraeli
Prime Minister of Great Britain (1868, 1874-1880)

When Statistics Seem to Lie

– They’re Answering a Different Question

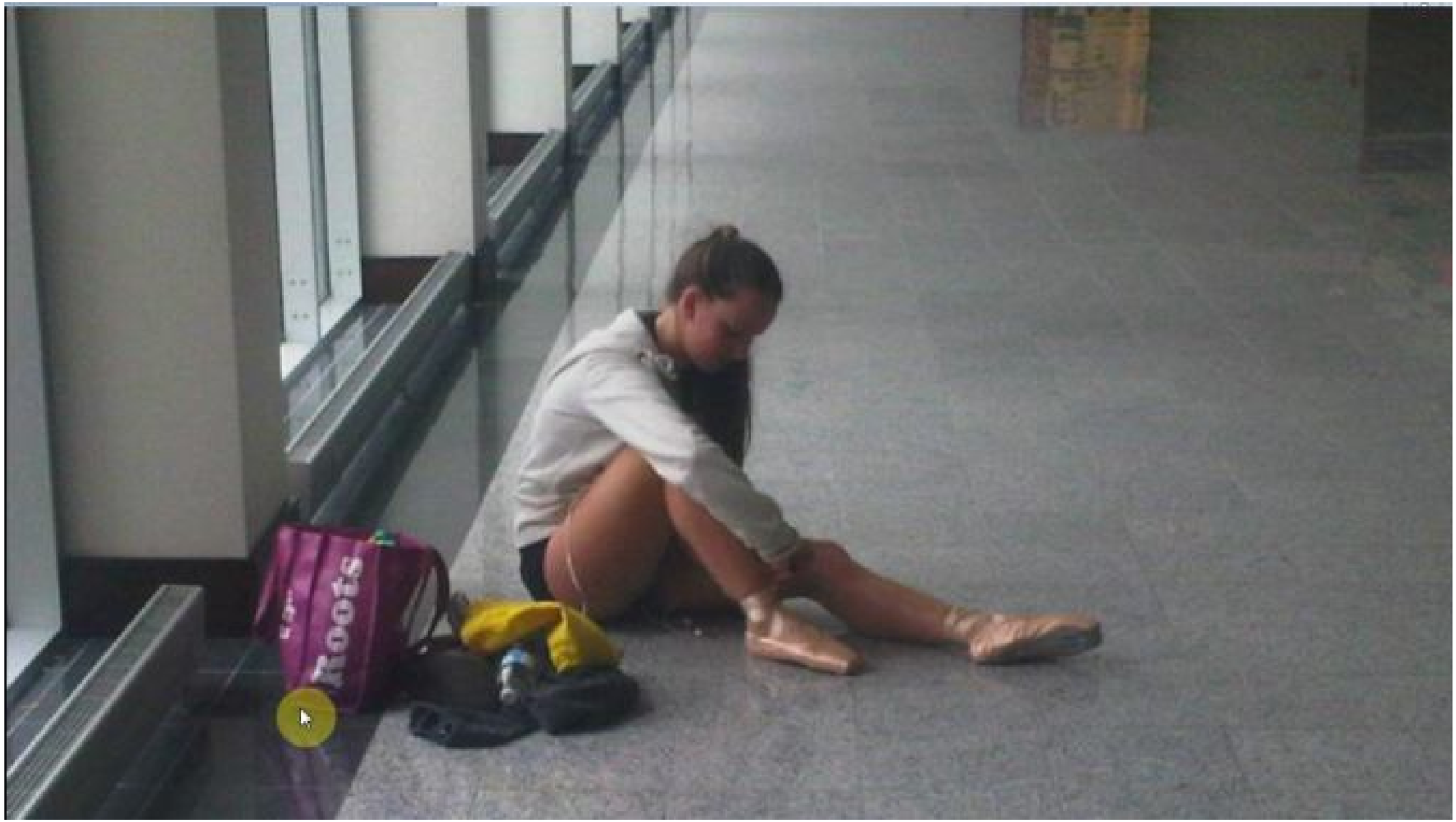
Georges Monette
York University

georges@yorku.ca


> STAR EXCLUSIVE
va keeps Ottawa keeps Ot
eviews drug reviews dru
vraps under wraps un
ssments of 151
d, will stay secret
Doctors alarmed that reassessment
> STAR EXCLUSIVE


GARDASIL®
Suspension injectable en seringue préremplie
1 dose, 0,5 ml - voie IM.


KAITLYN ARMSTRONG
Whitby, Ontario


Roots


LINDA MORIN
Laval, Quebec


MORIN

Science shows HPV vaccine has no dark side

To attribute rare devastating occurrences to a vaccine requires evidence of causation, which the Star didn't have in its article on Gardasil.

Given the power of HPV vaccine to prevent disease and death, a long Toronto Star article that appears to suggest that the HPV vaccine causes harm is troubling and disappointing, write Juliet Guichon and Dr. Rupert Kaul.

By: Juliet Guichon Dr. Rupert Kaul Published on Wed Feb 11 2015

The HPV vaccine was created to prevent an infection that causes cancer. That is pretty exciting. After all, Terry Fox's arduous marathon a day was to raise money for a cancer cure. Did he even imagine that we would have a vaccine to prevent cancer?

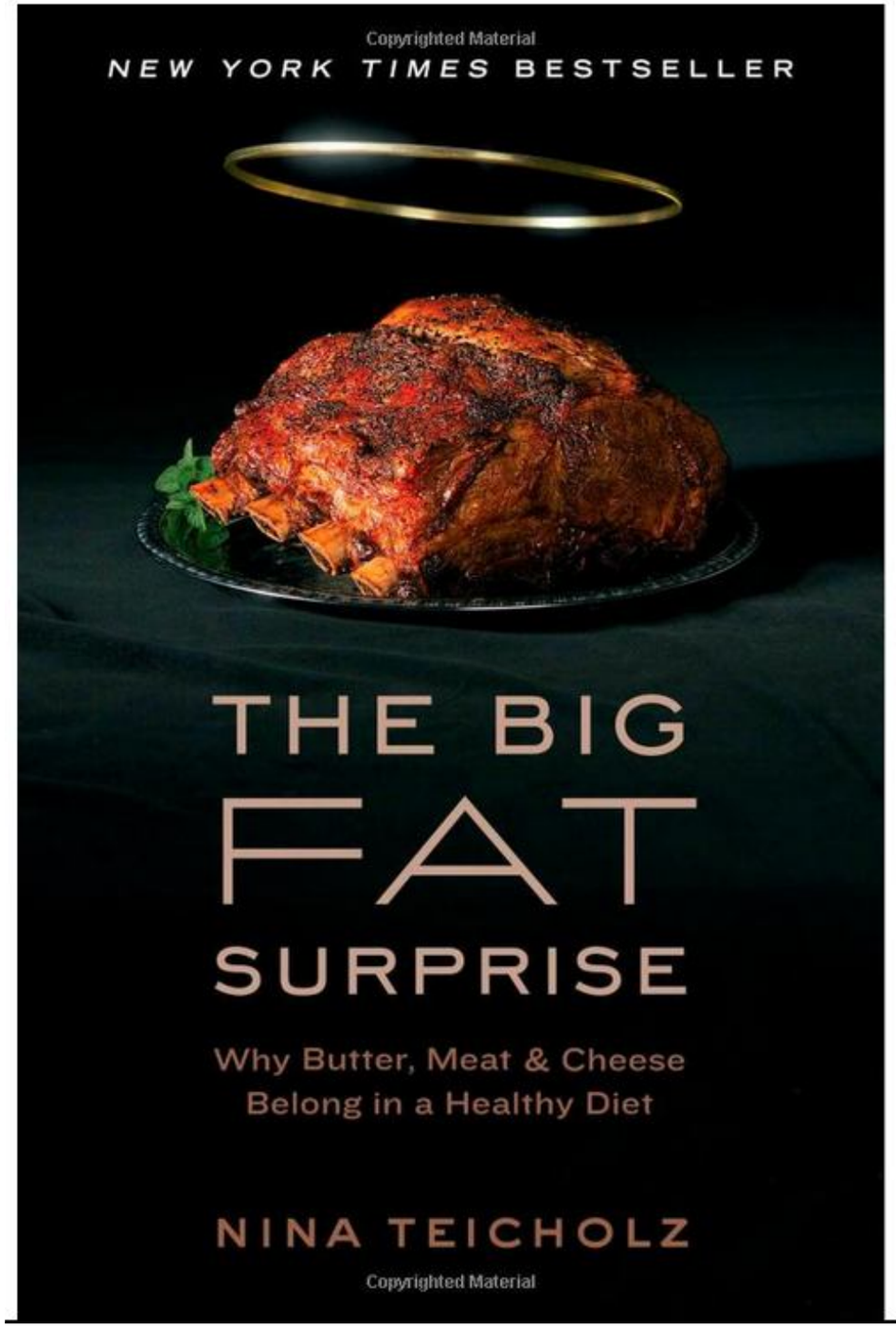
Given the power of HPV vaccine to prevent disease and death, a long Toronto Star article that appears to suggest that the HPV vaccine causes harm is troubling and disappointing. Although the article states in the fifth paragraph that "there is no conclusive evidence showing the vaccine caused a death or illness," its litany of horror stories and its innuendo give the incorrect impression that the vaccine caused the harm.

The Star story states that some people became sick and even died after being vaccinated against HPV infection. Yet, after HPV vaccination, some people might have won a major scholarship or the lottery. Does this mean the vaccine caused the award or the win? Hardly.

The fact that one event follows another does not mean that the first event caused the second — in scientific terms, correlation is not causation.

For example, the number of shark attacks and ice cream sales rise when the weather is hot. The confusion of correlation and causation here is funny because, of course, the shark attacks don't cause the ice cream sales increase. But in the case of the HPV vaccine, such confusion is not funny because HPV infection can have very serious consequences that the vaccine helps prevent.

The Star presented the stories of women who have suffered greatly. The article was engaging, dramatic and might have created fear. But study after study has shown that there is no causal link between the events the Star reported and the vaccine. About 169 million doses of the HPV vaccine have been administered worldwide. In any given large population, there will be illness and death. This is a statistical fact. To attribute rare devastating occurrences to a vaccine requires evidence of causation, of which the international scientific community and the Star article have none.



"Solid, well-reported science . . . Like a bloodhound, Teicholz tracks the process by which a hypothesis morphs into truth without the benefit of supporting data."

—*Kirkus Reviews* (starred review)

by Lara Goodrich Ezor June 5, 2014 4:55 AM

Butter Is NOT Back (And Other Truths About Saturated Fat)

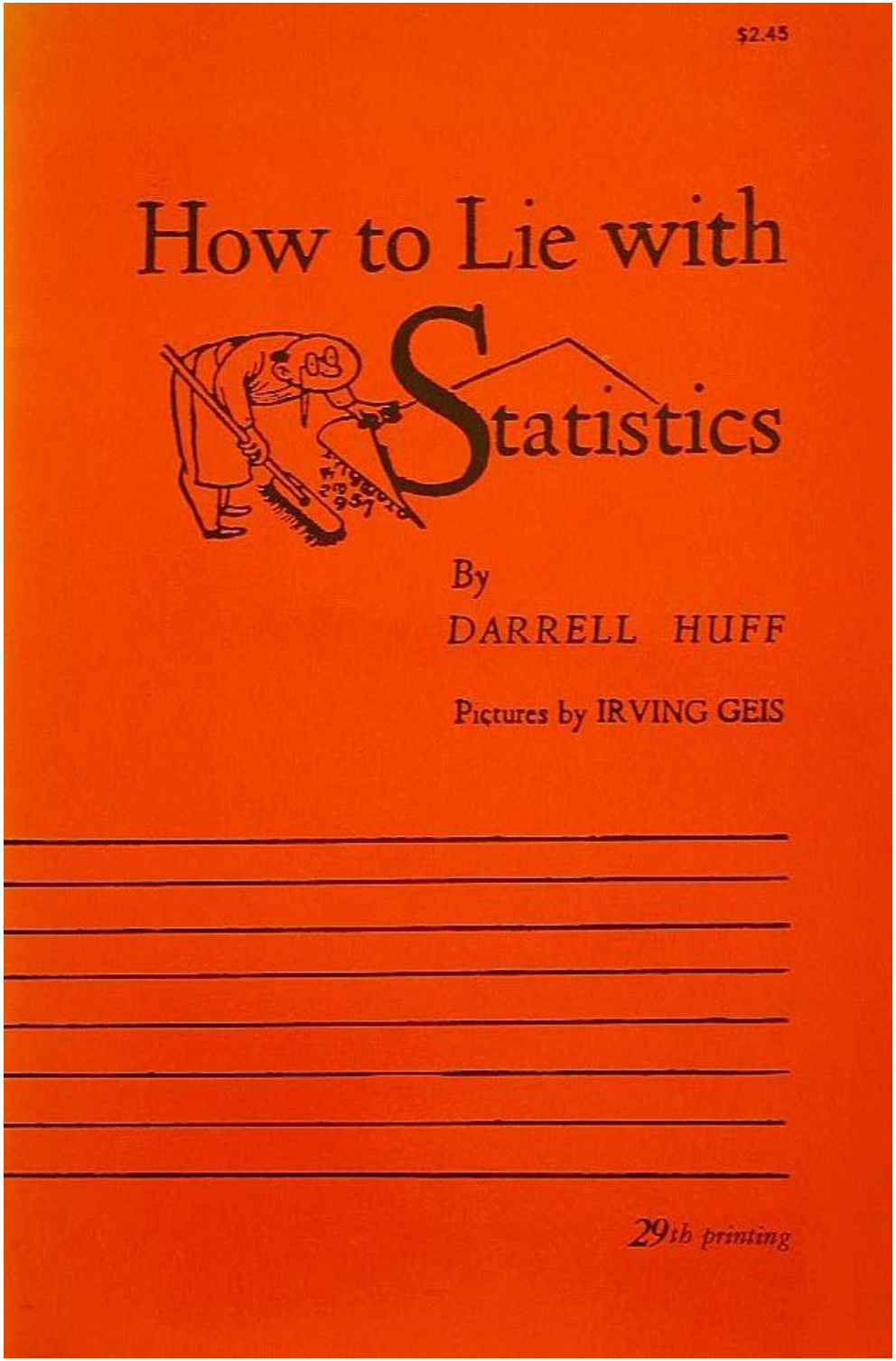
In March, *New York Times* writer and famous foodie Mark Bittman declared that "butter is back." His piece reported on the findings of a recent meta-analysis published in the *Annals of Internal Medicine* that questioned the long-standing link between saturated fat and coronary disease.

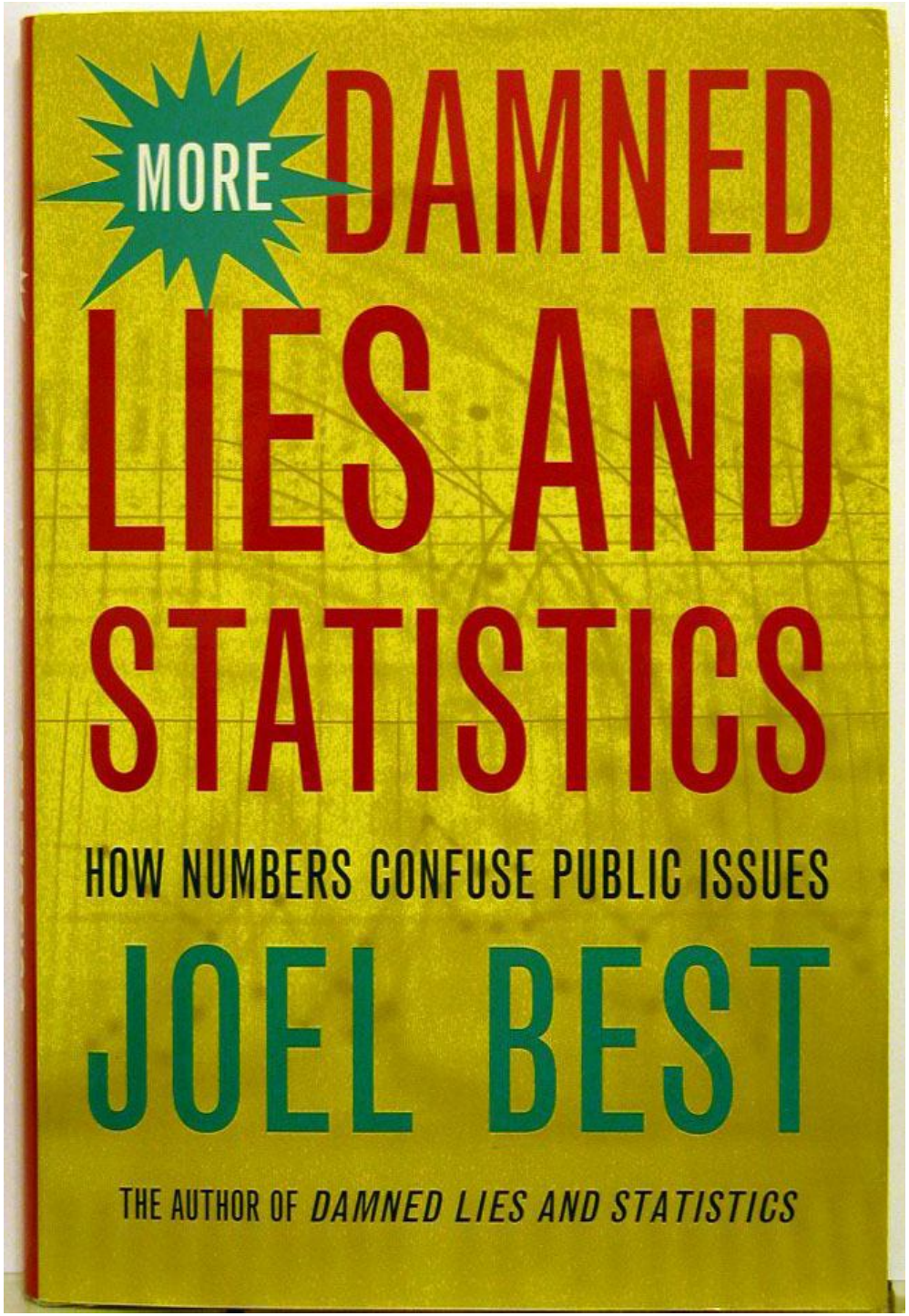
While Bittman celebrated the findings and told readers they could "go back to eating butter," nutrition and public health professionals have been quick to caution, "Not so fast!"

Dr. David Katz, Director of the Yale Prevention and Research Center, responded to the piece, pointing out Bittman's lack of qualifications for interpreting scientific studies and ultimately calling the writer "a potential danger to the public health."

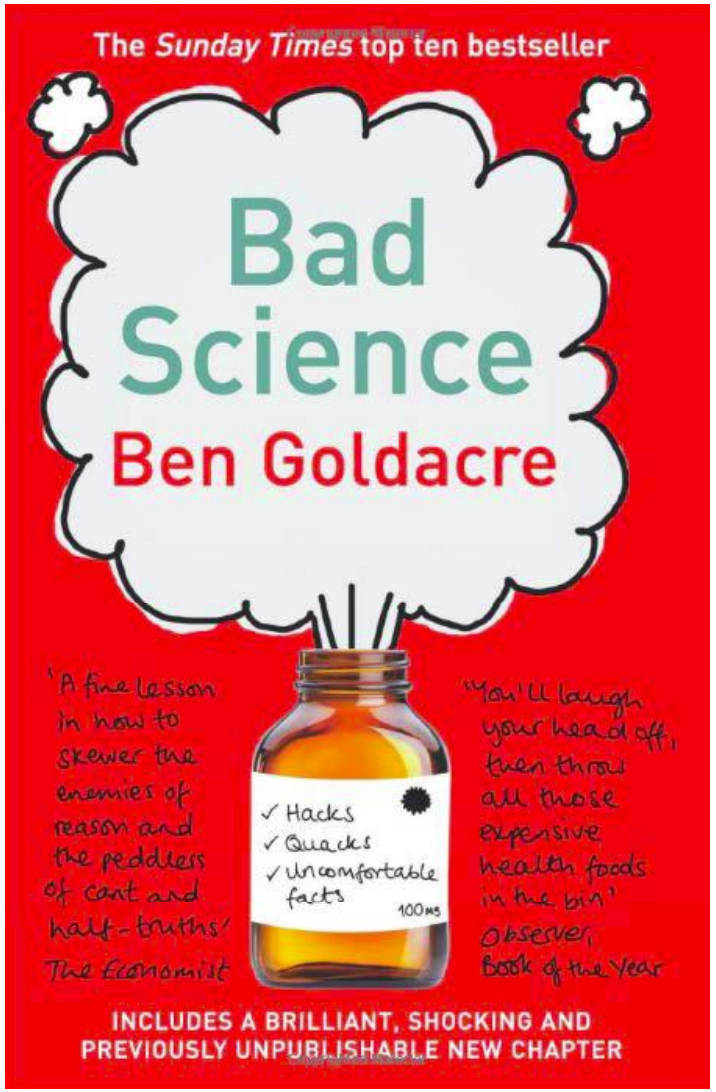
The Harvard School of Public Health put out a statement in the wake of the meta-analysis' publication calling its conclusions "seriously misleading," highlighting "many errors and omissions."

Best selling stats book of all times

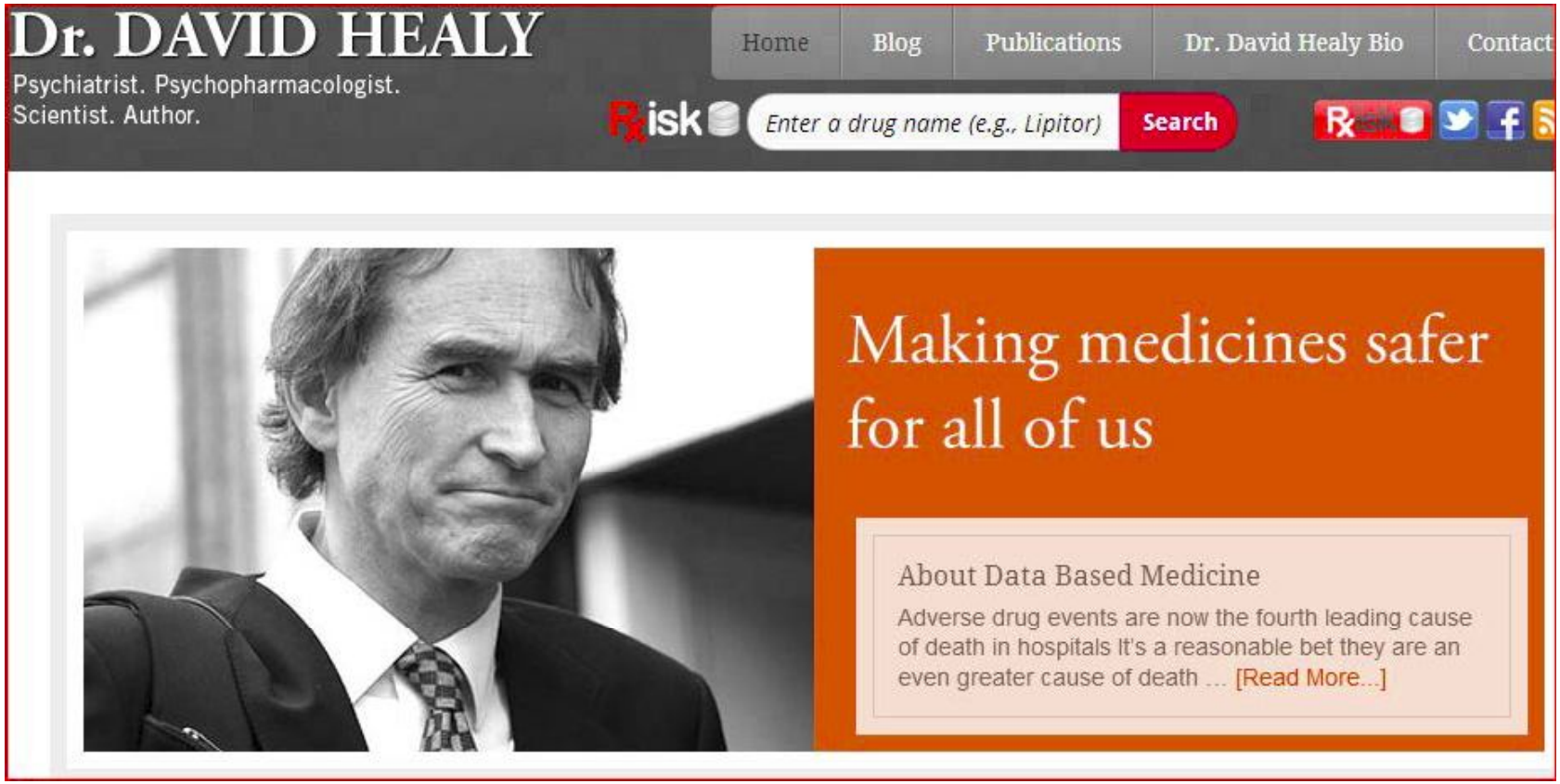


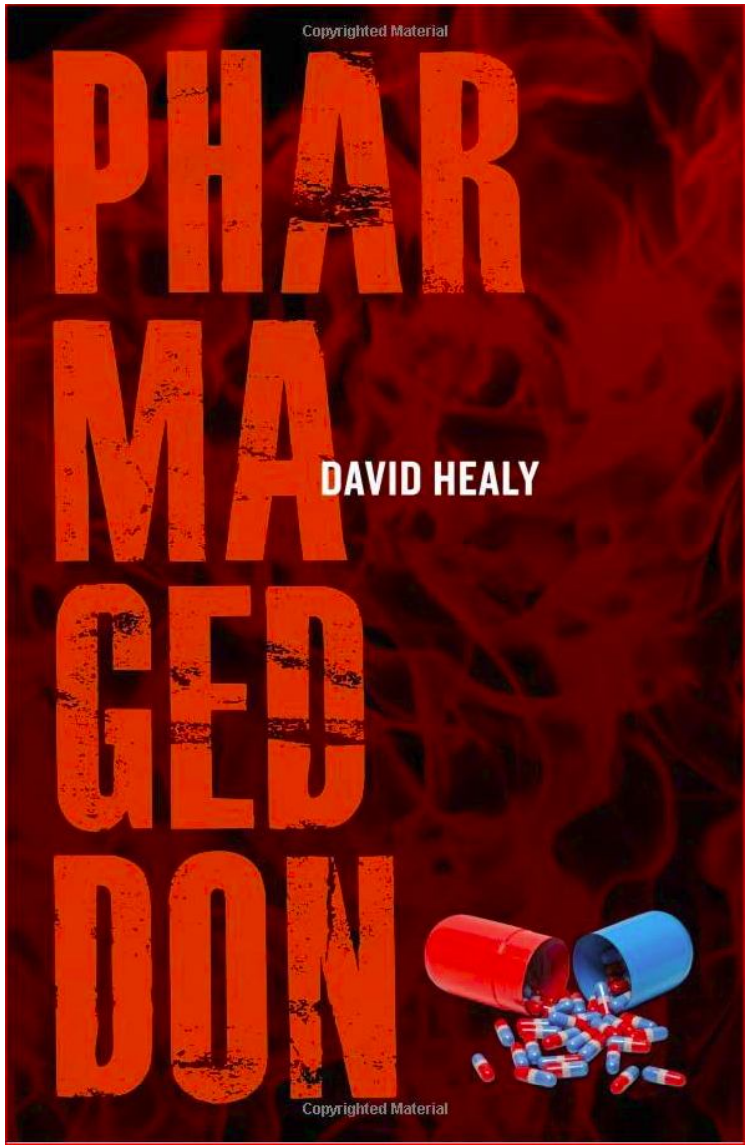

MORE
DAMNED
LIES AND
STATISTICS
HOW NUMBERS CONFUSE PUBLIC ISSUES
JOEL BEST
THE AUTHOR OF DAMNED LIES AND STATISTICS


Bad Pharma™
Ben Goldacre
Bestselling author of Bad Science
How drug companies mislead doctors and harm patients
364 pages
4th


The Sunday Times top ten bestseller
Bad Science
Ben Goldacre
'A fine lesson in how to skewer the enemies of reason and the peddlers of cant and half-truths' The Economist
✓ Hacks
✓ Quacks
✓ Uncomfortable facts
100ms
'You'll laugh your head off, then throw all those expensive health foods in the bin' Observer, Book of the Year
INCLUDES A BRILLIANT, SHOCKING AND PREVIOUSLY UNPUBLISHABLE NEW CHAPTER

Going further: David Healy (of CAMH fame):





David Healy takes Goldacre's argument one step further and questions whether relying Clinical Trials can give us the answers we need.

Statistical thinking will one day
be as necessary
for efficient citizenship
as the ability
to read and write.

– *H. G. Wells*

Misunderstand statistics?
Splitting hairs?
Does it really matter?

Misunderstand statistics?
Splitting hairs?
Does it really matter?

A few consequences?

- The global economic meltdown
- Wrongful murder convictions
- Delayed response to health effects of tobacco
- Poor health policies and treatment decisions

Meet the man whose big idea felled Wall Street

Math whiz proposed applying this statistical formula to credit risk, and financial meltdown ensued

Mar 18, 2009 04:30 AM

Comments on this story (102)

CATHAL KELLY
STAFF REPORTER

Note: This article has been edited to correct a previously published version.

Former University of Waterloo statistician David X. Li didn't burn down the American economy. He just supplied the matches.

University of Waterloo statistician David Li is shown in this handout photo, along with his statistical formula for modeling the behaviour of several correlated risks at once.

As economists and market watchers cast about for people to blame for the U.S. market meltdown, Li has surfaced as a scapegoat. Recently, *Wired* magazine ran an article on Li's work subtitled, "The Formula That Killed Wall Street."

The formula in question is the so-called Gaussian copula function. On the most basic level, the formula allows statisticians to model the behaviour of several correlated risks at once.

In a scholarly paper published in 2000, Li proposed the theorem be applied to credit risks, encompassing everything from bonds to mortgages. This particular copula was not new, but the financial application Li proposed for it was.

Disastrously, it was just simple enough for untrained financial analysts to use, but too complex for them to properly understand. It appeared to allow them to definitively determine risk, effectively eliminating it. The result was an orgy of misspending that sent the U.S. banking system over a cliff.

"To say David brought down the market is like blaming Einstein for Hiroshima," says Prof. Harry Panjer, Li's mentor at the University of Waterloo. "He wasn't in charge of the financial world. He just wrote an article."

It is easy to lie with statistics.
It is hard to tell the truth without it.

– *Andrejs Dunkels*

Pot use before 18 lowers IQ by 8 points

THERESA BOYLE
HEALTH REPORTER

Persistent, dependent use of marijuana before age 18 has been shown to cause lasting harm to a person's intelligence, attention and memory, according to a study in The Proceedings of the National Academy of Sciences of the U.S.

Among a long-range study cohort of more than 1,000 New Zealanders, individuals who started using cannabis in adolescence and used it for years afterward showed an average decline in IQ of eight points when their IQs were compared at ages 13 and 38. Quitting pot did not appear to reverse the loss either, said lead re-

Pot use before 18 lowers IQ by 8 points

THERESA BOYLE
HEALTH REPORTER

Persistent, dependent use of marijuana before age 18 has been shown to cause lasting harm to a person's intelligence, attention and memory, according to a study in The Proceedings of the National Academy of Sciences of the U.S.

Among a long-range study cohort of more than 1,000 New Zealanders, individuals who started using cannabis in adolescence and used it for years afterward showed an average decline in IQ of eight points when their IQs were compared at ages 13 and 38. Quitting pot did not appear to reverse the loss either, said lead re-

Don't forget to brush your teeth

Good oral health could lower risk of dementia

NATASJA SHERIFF
REUTERS

People who keep their teeth and gums healthy with regular brushing may have a lower risk of developing dementia later in life, according to a new study.

Researchers, who followed close to 5,500 elderly people over an 18-year period, found those who reported brushing their teeth less than once a day were up to 65 per cent more likely to develop dementia than those who brushed daily.

Not just global issues.
Also everyday decisions:

Does using cellphones cause brain cancer?

Plastic bottles? Are they poisonous?

- Controversy over Bisphenol-A bottles

New drugs: are they safe?

Will taking more Vitamin D help to prevent cancer?

Most of these issues boil down to asking:

Will X cause Y?

Why can’t the experts agree?

How do I make a wise decision for myself?

Should I or Shouldn't I do X?

Does doing X cause Y?

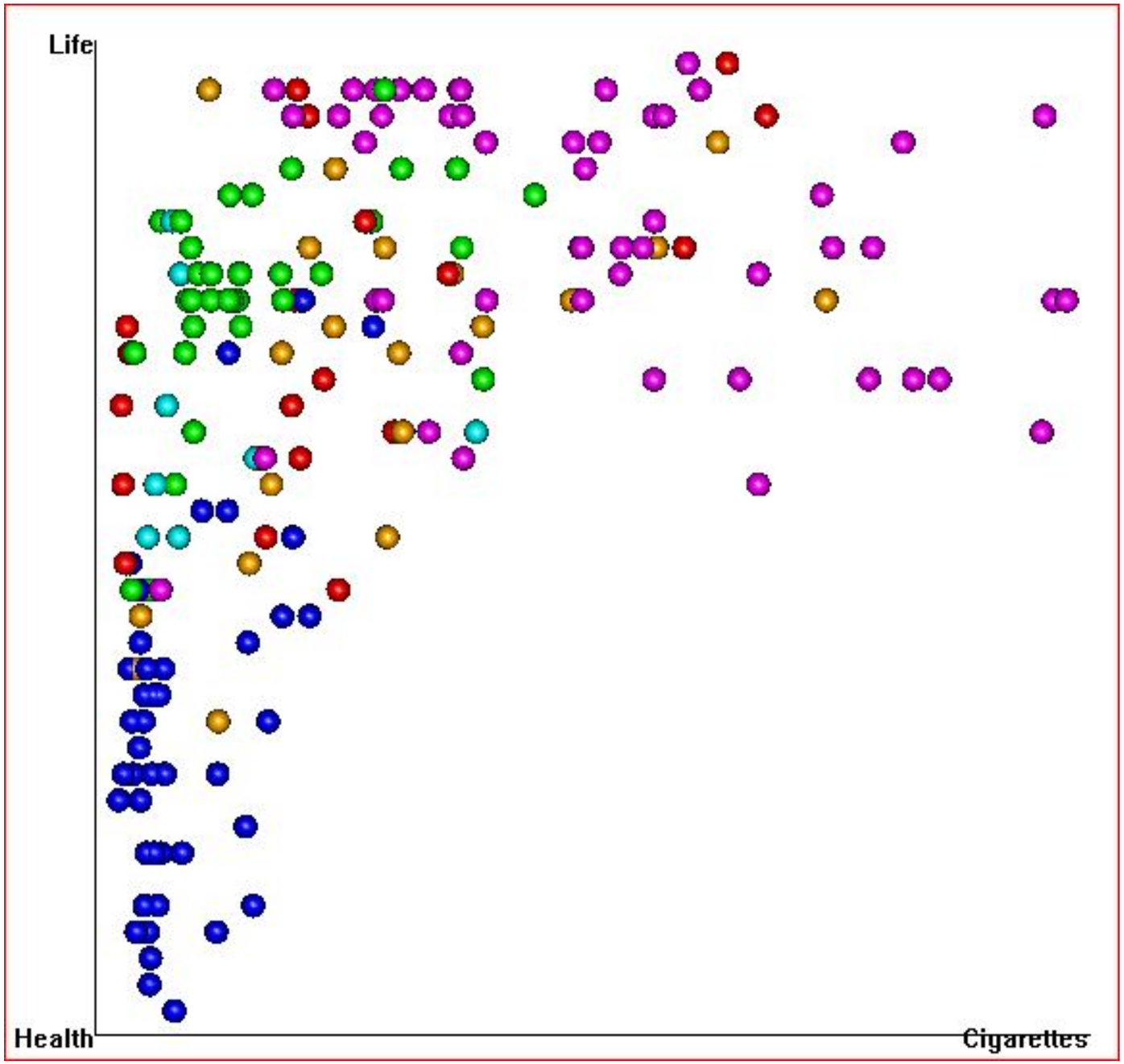
Answering an important question:

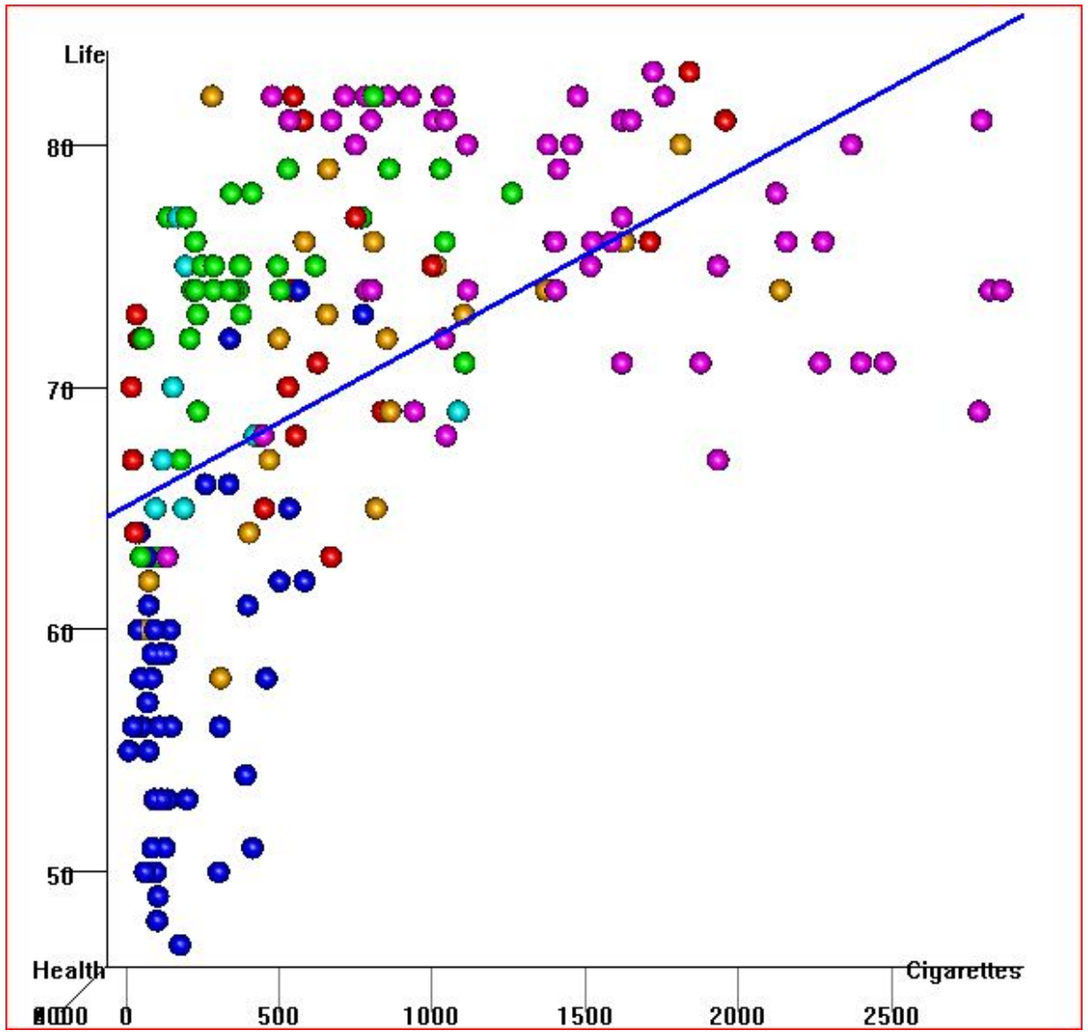
Just how harmful is smoking anyways?

Use data for an ‘evidence-based’ answer:

We can go to the web (e.g. Gapminder.org) to get data on
Smoking and on **Life Expectancy**
from most countries in the world

We’ll see just how much smoking is bad for your health by looking at
the **relationship** between **Smoking** and **Life Expectancy**


Life
Health
Cigarettes



Coefficients	Estimate	Std.Error	DF	t-value	p-value
(Intercept)	65.075840	0.855974	183	76.025515	<.00001
Cigarettes	0.006915	0.000855	183	8.090493	<.00001

Coefficients	Estimate	Std.Error	DF	t-value	p-value
(Intercept)	65.075840	0.855974	183	76.025515	<.00001
Cigarettes	0.006915	0.000855	183	8.090493	<.00001

What does this actually mean?

```
Coefficients   Estimate Std.Error  DF   t-value p-value
  (Intercept) 65.075840  0.855974 183 76.025515 <.00001
  Cigarettes   0.006915  0.000855 183  8.090493 <.00001
```

What does this actually mean?

One extra **cigarette per year** adds

0.006915 years to your life,

```
Coefficients   Estimate Std.Error  DF   t-value p-value
  (Intercept) 65.075840  0.855974 183 76.025515 <.00001
  Cigarettes   0.006915  0.000855 183  8.090493 <.00001
```

What does this actually mean?

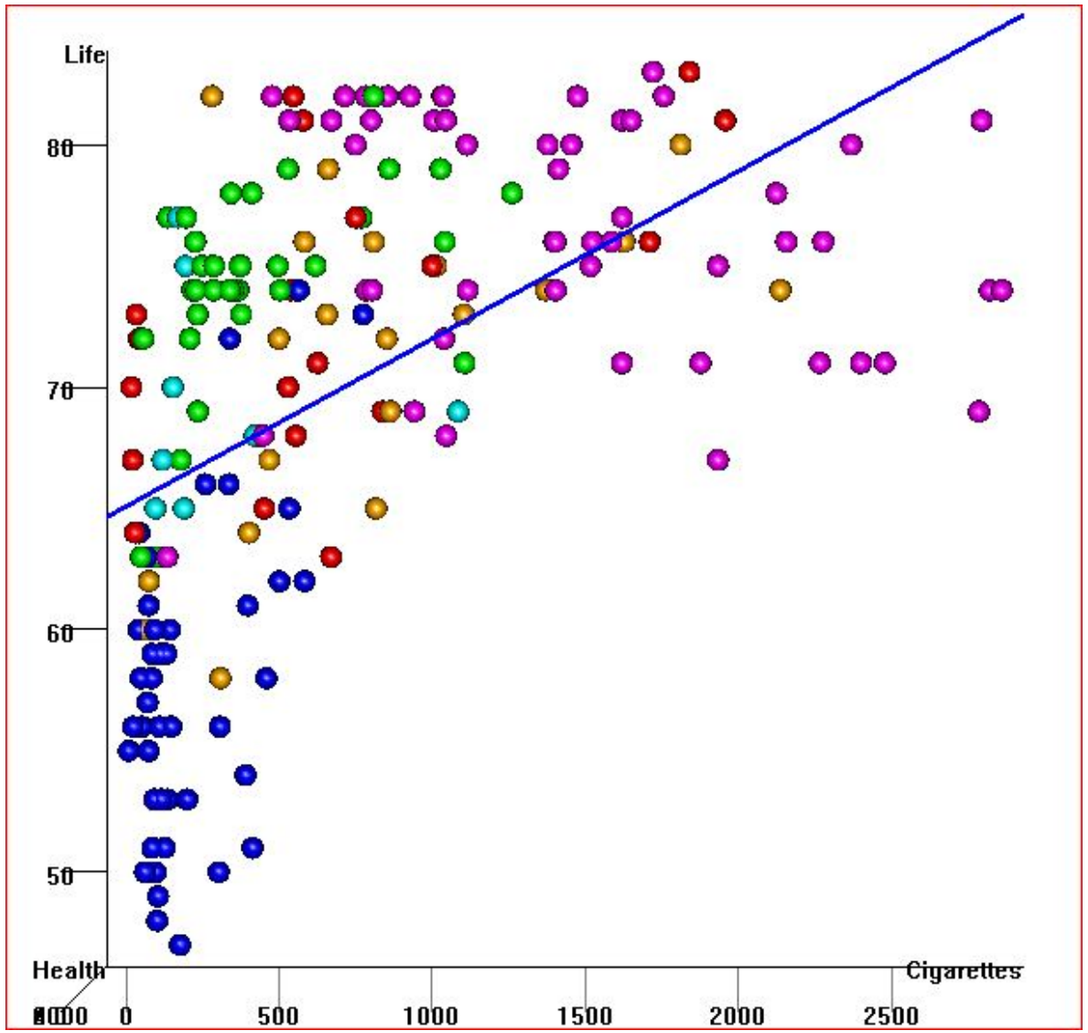
One extra **cigarette per year** adds

0.006915 years to your life,

Not very impressive but in better units:

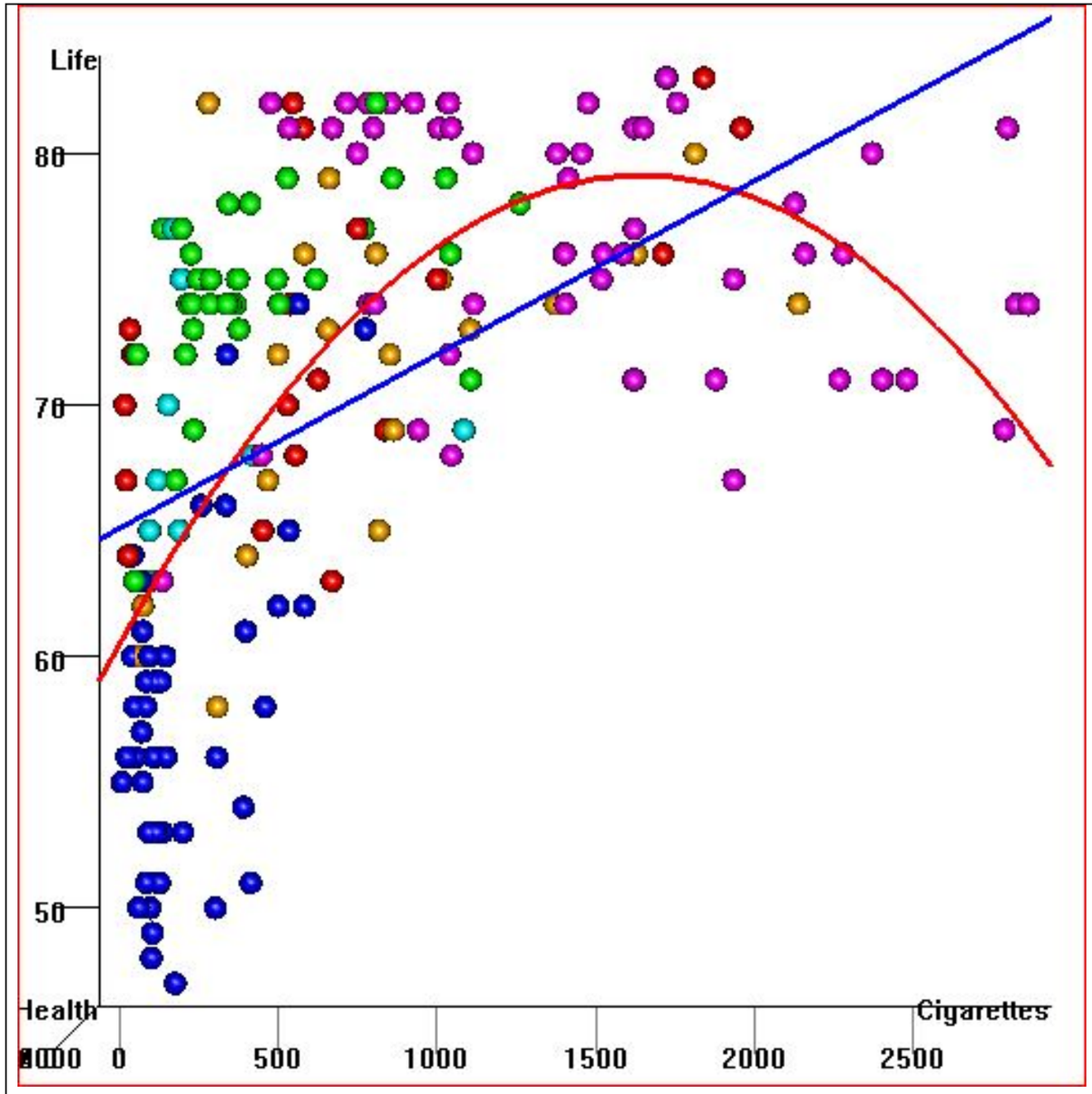
All it takes is 4 cigarettes a day

to add 10 years to your life


Life
80
70
60
50
Health
Cigarettes
0
500
1000
1500
2000
2500

A good statistician would tell you
that this is ridiculous.

There’s obvious curvature in the relationship

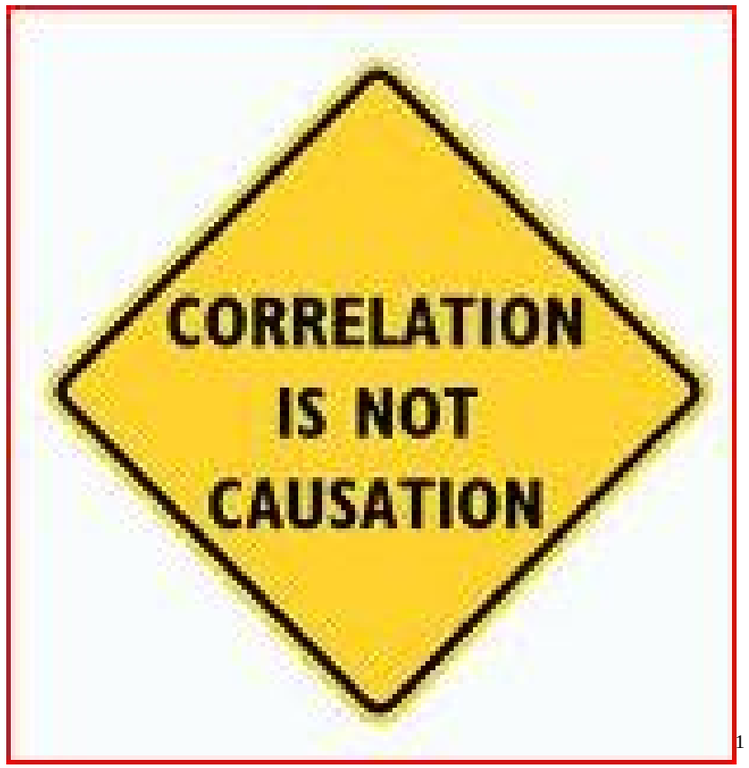


Fitting a quadratic model and maximizing the quadratic shows that

4.495 cigarettes/day

is actually optimal

What’s the problem?



[1]

[1] Adapted from a sign by Edward Tufte

Maybe it isn’t smoking that’s responsible for higher life expectancies.

Maybe it's something else –
a **CONFOUNDING VARIABLE**
(also called a "**LURKING VARIABLE**" or "**LURKING FACTOR**")
that causes **BOTH**
higher life expectancies
and higher rates of smoking.

1) $X \Rightarrow Y$

1) $X \Rightarrow Y$ i.e. $\Delta\uparrow X \Rightarrow \uparrow E(Y)$

1) $X \Rightarrow Y$ i.e. $\Delta\uparrow X \Rightarrow \uparrow E(Y)$ Maybe a) Directly $X \Rightarrow Y$

1) $X \Rightarrow Y$ i.e. $\Delta\uparrow X \Rightarrow \uparrow E(Y)$

Maybe

a) Directly $X \Rightarrow Y$

b) Through mediating factor(s)

$X \Rightarrow M_1 \Rightarrow Y$

$X \Rightarrow M_2 \Rightarrow Y$

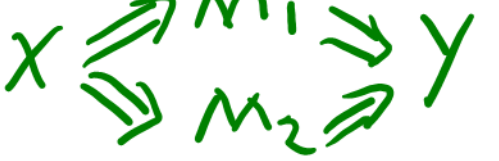
1) $X \Rightarrow Y$ i.e. $\Delta\uparrow X \Rightarrow \uparrow E(Y)$

2) $Y \Rightarrow X$

Maybe

a) Directly $X \Rightarrow Y$

b) Through mediating factor(s)



1) $X \Rightarrow Y$ i.e. $\Delta\uparrow X \Rightarrow \uparrow E(Y)$

2) $Y \Rightarrow X$

3) Z

Maybe

a) Directly $X \Rightarrow Y$

b) Through mediating factor(s)

1) $X \Rightarrow Y$ i.e. $\Delta\uparrow X \Rightarrow \uparrow E(Y)$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$

Maybe

a) Directly $X \Rightarrow Y$

b) Through mediating factor(s)

$X \Rightarrow M_1 \Rightarrow Y$

$X \Rightarrow M_2 \Rightarrow Y$

1) $X \Rightarrow Y$ i.e. $\Delta\uparrow X \Rightarrow \uparrow E(Y)$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$

Maybe

a) Directly $X \Rightarrow Y$

b) Through mediating factor(s)

$X \Rightarrow M_1 \Rightarrow Y$

$X \Rightarrow M_2 \Rightarrow Y$

1) $X \Rightarrow Y$ i.e. $\Delta\uparrow X \Rightarrow \uparrow E(Y)$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$ Confounding factor(s)

Maybe

a) Directly $X \Rightarrow Y$

b) Through mediating factor(s)

$X \Rightarrow M_1 \Rightarrow Y$

$X \Rightarrow M_2 \Rightarrow Y$

1) $X \Rightarrow Y$ i.e. $\Delta \uparrow X \Rightarrow \uparrow E(Y)$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$ Confounding factor(s)

a) Z known + measurable

Maybe

a) Directly $X \Rightarrow Y$

b) Through mediating factor(s)

$X \Rightarrow M_1 \Rightarrow Y$

$X \Rightarrow M_2 \Rightarrow Y$

1) $X \Rightarrow Y$ i.e. $\Delta \uparrow X \Rightarrow \uparrow E(Y)$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$

Confounding factor(s)

a) Z known + measurable — can control with multiple regression

Maybe

a) Directly $X \Rightarrow Y$

b) Through mediating factor(s)

$X \Rightarrow M_1 \Rightarrow Y$

$X \Rightarrow M_2 \Rightarrow Y$

1) $X \Rightarrow Y$ i.e. $\Delta\uparrow X \Rightarrow \uparrow E(Y)$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$ Confounding factor(s)

a) Z known + measurable

b) Z " but hard to measure

Maybe

a) Directly $X \Rightarrow Y$

b) Through mediating factor(s)

$X \Rightarrow M_1 \Rightarrow Y$

$X \Rightarrow M_2 \Rightarrow Y$

1) $X \Rightarrow Y$ i.e. $\Delta \uparrow X \Rightarrow \uparrow E(Y)$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$ — Confounding factor(s)

a) Z known & measurable

b) Z " but hard to measure → adjust for measurement error – SEMs

Maybe

a) Directly $X \Rightarrow Y$

b) Through mediating factor(s)

$X \Rightarrow M_1 \Rightarrow Y$

$X \Rightarrow M_2 \Rightarrow Y$

1) $X \Rightarrow Y$ i.e. $\Delta\uparrow X \Rightarrow \uparrow E(Y)$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$ — Confounding factor(s)

a) Z known + measurable

b) Z " but hard to measure

c) Z unknown

Maybe

a) Directly $X \Rightarrow Y$

b) Through mediating factor(s)

$X \Rightarrow M_1 \Rightarrow Y$

$X \Rightarrow M_2 \Rightarrow Y$

1) $X \Rightarrow Y$ i.e. $\Delta\uparrow X \Rightarrow \uparrow E(Y)$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$ Confounding factor(s)

a) Z known + measurable

b) Z " but hard to measure

c) Z unknown

d) There are clusters in which Z is constant

Maybe

a) Directly $X \Rightarrow Y$

b) Through mediating factor(s)

$X \Rightarrow M_1 \Rightarrow Y$

$X \Rightarrow M_2 \Rightarrow Y$

1) $X \Rightarrow Y$ i.e. $\Delta\uparrow X \Rightarrow \uparrow E(Y)$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$ — Confounding factor(s)

Maybe

a) Directly $X \Rightarrow Y$

b) Through mediating factor(s)

$X \Rightarrow M_1 \Rightarrow Y$

$X \Rightarrow M_2 \Rightarrow Y$

a) Z known + measurable

b) Z " but hard to measure

c) Z unknown

d) There are clusters in which Z is constant → use longitudinal or nested data

1) $X \Rightarrow Y$ i.e. $\Delta\uparrow X \Rightarrow \uparrow E(Y)$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$ — Confounding factor(s)

   a) Z known + measurable
   b) Z " but hard to measure
   c) Z unknown
   d) There are clusters in which Z is constant

4) Chance

Maybe
a) Directly $X \Rightarrow Y$
b) Through mediating factor(s)

$X \Rightarrow M_1 \Rightarrow Y$
$X \Rightarrow M_2 \Rightarrow Y$

1) $X \Rightarrow Y$ i.e. $\Delta\uparrow X \Rightarrow \uparrow E(Y)$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$ Confounding factor(s)

a) Z known + measurable

b) Z " but hard to measure

c) Z unknown

d) There are clusters in which Z is constant

4) Chance

5) Selection

Maybe

a) Directly $X \Rightarrow Y$

b) Through mediating factor(s)

$X \Rightarrow M_1 \Rightarrow Y$

$X \Rightarrow M_2 \Rightarrow Y$

1) $X \Rightarrow Y$ i.e. $\Delta\uparrow X \Rightarrow \uparrow E(Y)$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$ — Confounding factor(s)

   a) Z known + measurable
   b) Z " but hard to measure
   c) Z unknown
   d) There are clusters in which Z is constant

4) Chance

5) Selection

Maybe
a) Directly $X \Rightarrow Y$
b) Through mediating factor(s)

$X \Rightarrow M_1 \Rightarrow Y$
$X \Rightarrow M_2 \Rightarrow Y$

1) $X \Rightarrow Y$ i.e. $\Delta\uparrow X \Rightarrow \uparrow E(Y)$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$ — Confounding factor(s)

a) Z known + measurable

b) Z " but hard to measure

c) Z unknown

d) There are clusters in which Z is constant

4) Chance

5) Selection

Maybe

a) Directly $X \Rightarrow Y$

b) Through mediating factor(s)

$X \Rightarrow M_1 \Rightarrow Y$

$X \Rightarrow M_2 \Rightarrow Y$

- To conclude that $X \Rightarrow Y$ we need to be willing to reject the other possibilities.

1) $X \Rightarrow Y$ i.e. $\Delta\uparrow X \Rightarrow \uparrow E(Y)$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$ — Confounding factor(s)

a) Z known & measurable

b) Z " but hard to measure

c) Z unknown

d) There are clusters in which Z is constant

4) Chance

5) Selection

Maybe

a) Directly $X \Rightarrow Y$

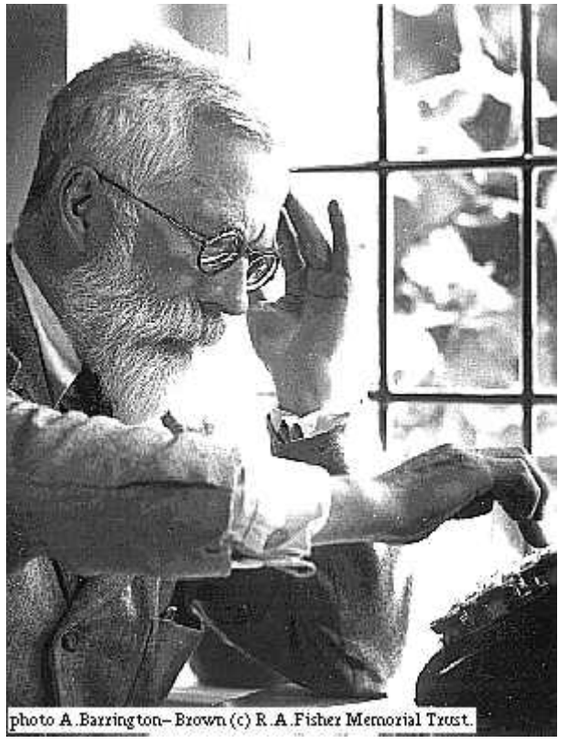
b) Through mediating factor(s)

$X \Rightarrow M_1 \Rightarrow Y$

$X \Rightarrow M_2 \Rightarrow Y$

- To conclude that $X \Rightarrow Y$ we need to be willing to reject the other possibilities.
- Ordinary statistical analysis only helps with #4 via p-value.

R. A. Fisher's brilliant solution (~1920):



Randomized Experiment

using **Random Assignment** to treatments (levels of the X variable)

To avoid the possibility that some factor other than smoking is responsible for the difference in health:

Toss a coin to choose who gets to smoke and who doesn't

Observe for many years and then compare smokers and non-smokers

If there's a difference between the two groups either smoking that's responsible **OR** it's due to something else ***BY CHANCE – which we can measure***

What can it mean if X is correlated (associated) with Y in a sample?

1) $X \Rightarrow Y$ i.e. $\Delta\uparrow X \Rightarrow \uparrow E(Y)$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$ — Confounding factor(s)

Maybe

a) Directly $X \Rightarrow Y$

b) Through mediating factor(s)

$X \Rightarrow M_1 \Rightarrow Y$

$X \Rightarrow M_2 \Rightarrow Y$

a) Z known + measurable

b) Z " but hard to measure

c) Z unknown

d) There are clusters in which Z is constant

4) Chance

5) Selection

- To conclude that $X \Rightarrow Y$ we need to be willing to reject the other possibilities.
- Ordinary statistical analysis only helps with #4 via p-value.

What can it mean if X is correlated (associated) with Y in a sample?

1) $X \Rightarrow Y$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$

   a) Z known + measurable
   b) Z " but hard to measure
   c) Z unknown
   d) There are clusters in which Z is constant

4) Chance

5) Selection

What can it mean if X is correlated (associated) with Y in a sample?

OBSERVATIONAL DATA

1) $X \Rightarrow Y$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$

a) Z known + measurable

b) Z " but hard to measure

c) Z unknown

d) There are clusters in which Z is constant

4) Chance

5) Selection

What can it mean if X is correlated (associated) with Y in a sample?

OBSERVATIONAL DATA

EXPERIMENTAL DATA

1) $X \Rightarrow Y$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$

   a) Z known + measurable
   b) Z " but hard to measure
   c) Z unknown
   d) There are clusters in which Z is constant

4) Chance

5) Selection

What can it mean if X is correlated (associated) with Y in a sample?

OBSERVATIONAL DATA

1) $X \Rightarrow Y$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$

   a) Z known + measurable
   b) Z " but hard to measure
   c) Z unknown
   d) There are clusters in which Z is constant

4) Chance

5) Selection

EXPERIMENTAL DATA

1)

What can it mean if X is correlated (associated) with Y in a sample?

OBSERVATIONAL DATA

1) $X \Rightarrow Y$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$

   a) Z known + measurable
   b) Z " but hard to measure
   c) Z unknown
   d) There are clusters in which Z is constant

4) Chance

5) Selection

EXPERIMENTAL DATA

1) ✓

What can it mean if X is correlated (associated) with Y in a sample?

OBSERVATIONAL DATA

1) $X \Rightarrow Y$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$

   a) Z known + measurable
   b) Z " but hard to measure
   c) Z unknown
   d) There are clusters in which Z is constant

4) Chance

5) Selection

EXPERIMENTAL DATA

1) ✓

2)

What can it mean if X is correlated (associated) with Y in a sample?

OBSERVATIONAL DATA

1) $X \Rightarrow Y$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$

   a) Z known + measurable
   b) Z " but hard to measure
   c) Z unknown
   d) There are clusters in which Z is constant

4) Chance

5) Selection

EXPERIMENTAL DATA

1) ✓

2) ?

What can it mean if X is correlated (associated) with Y in a sample?

OBSERVATIONAL DATA

1) $X \Rightarrow Y$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$

   a) Z known + measurable
   b) Z " but hard to measure
   c) Z unknown
   d) There are clusters in which Z is constant

4) Chance

5) Selection

EXPERIMENTAL DATA

1) ✓

2) X "caused" by coin toss

What can it mean if X is correlated (associated) with Y in a sample?

OBSERVATIONAL DATA

1) $X \Rightarrow Y$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$

   a) Z known + measurable
   b) Z " but hard to measure
   c) Z unknown
   d) There are clusters in which Z is constant

4) Chance

5) Selection

EXPERIMENTAL DATA

1) ✓

2) ~~X~~ X "caused" by coin toss

What can it mean if X is correlated (associated) with Y in a sample?

OBSERVATIONAL DATA

1) $X \Rightarrow Y$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$

   a) Z known + measurable
   b) Z " but hard to measure
   c) Z unknown
   d) There are clusters in which Z is constant

4) Chance

5) Selection

EXPERIMENTAL DATA

1) ✓

2) ~~X~~ X "caused" by coin toss

3) ?

What can it mean if X is correlated (associated) with Y in a sample?

OBSERVATIONAL DATA

1) $X \Rightarrow Y$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$

   a) Z known + measurable
   b) Z " but hard to measure
   c) Z unknown
   d) There are clusters in which Z is constant

4) Chance

5) Selection

EXPERIMENTAL DATA

1) ✓

2) ~~X~~ X "caused" by coin toss

3) } by chance

What can it mean if X is correlated (associated) with Y in a sample?

OBSERVATIONAL DATA

1) $X \Rightarrow Y$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$

   a) Z known + measurable
   b) Z " but hard to measure
   c) Z unknown
   d) There are clusters in which Z is constant

4) Chance

5) Selection

EXPERIMENTAL DATA

1) ✓

2) ~~✓~~ X "caused" by coin toss

3) } by chance

4) ✓

What can it mean if X is correlated (associated) with Y in a sample?

OBSERVATIONAL DATA

1) $X \Rightarrow Y$

2) $Y \Rightarrow X$

3) $Z \Rightarrow X$, $Z \Rightarrow Y$

   a) Z known + measurable
   b) Z " but hard to measure
   c) Z unknown
   d) There are clusters in which Z is constant

4) Chance

5) Selection

EXPERIMENTAL DATA

1) ✓

2) ~~✓~~ X "caused" by coin toss

3) } by chance

4) ✓

5) ✓

Should we only use experimental data to determine whether X causes Y?

Problems with experimental data:

- too costly
- too risky
- too long
- subjects who are willing and available may not be typical of target population
- observational data already on hand so let's use it
- won't give an answer until it's too late
- experimental situation not realistic
- we can only tell whether **assignment to treatment groups** makes a difference. What if subjects don’t comply?

For example: clinical trials are used to assess the **effectivenes** of drugs but not useful to discover possible rare side-effects. These need to be monitored with observational data when the drug is being used.

"Second best" method for causal inference:

Use observational data with care

How?

Use *observational data* and try to control for the possible effects of a confounding factor(s) by measuring it and

1) Analyzing each *stratum* with similar values for the confounding factor(s). This is called *stratification*.

OR

2) Building a statistical model in that includes the confounding factor(s) and using *multiple regression*.

OR

3) Use new advanced methods: propensity score matching, discontinuity models, etc.

This are no perfect solutions and they all require judgment to assess studies based on these methods:

Problems:

1) The confounding factor may be known but may be measured with error so that it is not fully controlled.

2) Some important confounding factors might not be known.

Note that these are NOT problems for randomized experiments.

Understanding the problem:

The fundamental 2 x 2 table of statistics			
Questions			

Understanding the problem:

The fundamental 2 x 2 table of statistics			
Questions	**Causal** what would happen if …?		
	Predictive passive guessing		

Understanding the problem:

The fundamental 2 x 2 table of statistics		**Data**	
		Experimental random assignment to treatments (X)	**Observational** X is not controlled
Questions	**Causal** what would happen if …?		
	Predictive passive guessing		

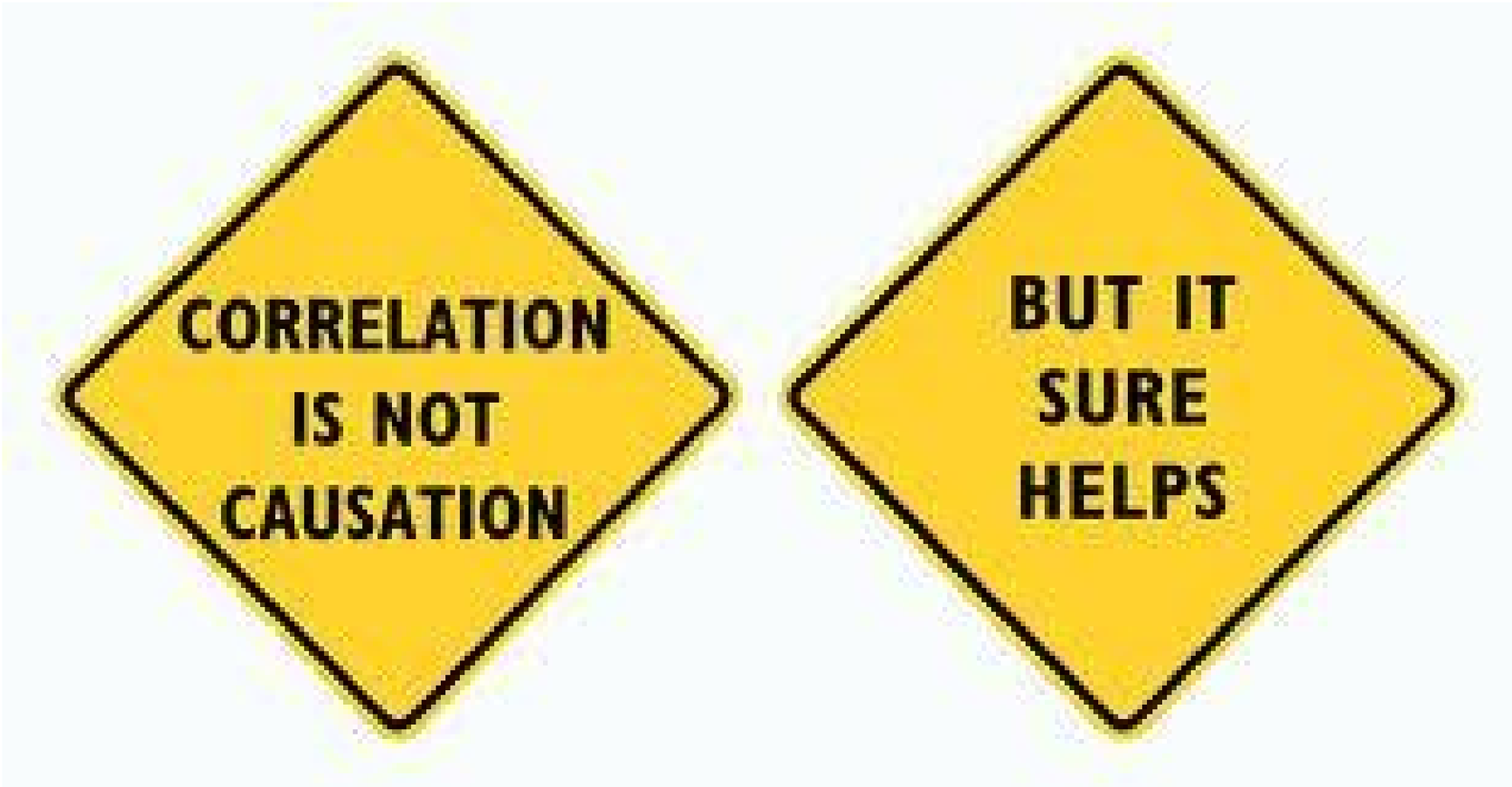
Understanding the problem:

The fundamental 2 x 2 table of statistics		**Data**	
		Experimental random assignment to treatments (X)	**Observational** X is not controlled
Questions	**Causal** what would happen if …?	**Ideal** where Fisher wants to be	
	Predictive passive guessing		**Ideal** for prediction under the same conditions as those observed

Understanding the problem:

The fundamental 2 x 2 table of statistics		**Data**	
		Experimental random assignment to treatments (X)	**Observational** X is not controlled
Questions	**Causal** what would happen if …?	**Ideal** where Fisher wants to be	**Where most of the difficult questions are**
	Predictive passive guessing	Hardly ever	**Ideal** for prediction under the same conditions as those observed

Hints of causal effects based on correlations (observational data) are everywhere:



How should we react to them?

(how would we like our students to react to them)

How can we do better than Fisher?

Should we even try?

Recent example in the news:

People who use sunscreen lotion have a higher risk of skin cancer than people who don’t

Should I stop using SSL?

How can we make wise decisions when faced with this kind of information?

The solution to the problem involves
asking questions
more than finding answers!

What question do we want to ask?
Is the question causal or predictive?

What kind of data do we have?

How were people assigned randomly
to use more or less SSL?

If the answer is yes, then we go on to ask more questions: Were the subjects like me? Did they comply with the random assignment?

If the answer is ‘not randomly’ then we need to think of possible confounding factors.

Understanding these issues is important for simple everyday questions.

But also for very large questions

Conjectures::

1. Most scientific and social controversies subsist on conflicting interpretations of evidence

2. Most conflicting interpretations of evidence are rooted in difficulties inferring causality from observational data

Caution:

Taking a hard line “**correlation is not causation**”
may be as problematic as seeing causation in every correlation.

Caution:

Taking a hard line "**correlation is not causation**"
may be as problematic as seeing causation in every correlation.

For many important issues, we only have observational data.

This is a major challenge for modern Statistics and for the interpretation of scientific evidence.

We need to find a balance between extreme skepticism and extreme gullibility.