The Fundamental 2 x 2 Table of Statistics

When Statistics Seem to Lie They are Usually Just Answering a Different Question

Georges Monette

random@yorku.ca

There are three kinds of lies:

lies, damned lies and statistics

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

Ottawa keeps drug reviews under wraps

sments of 151
d, will stay secret
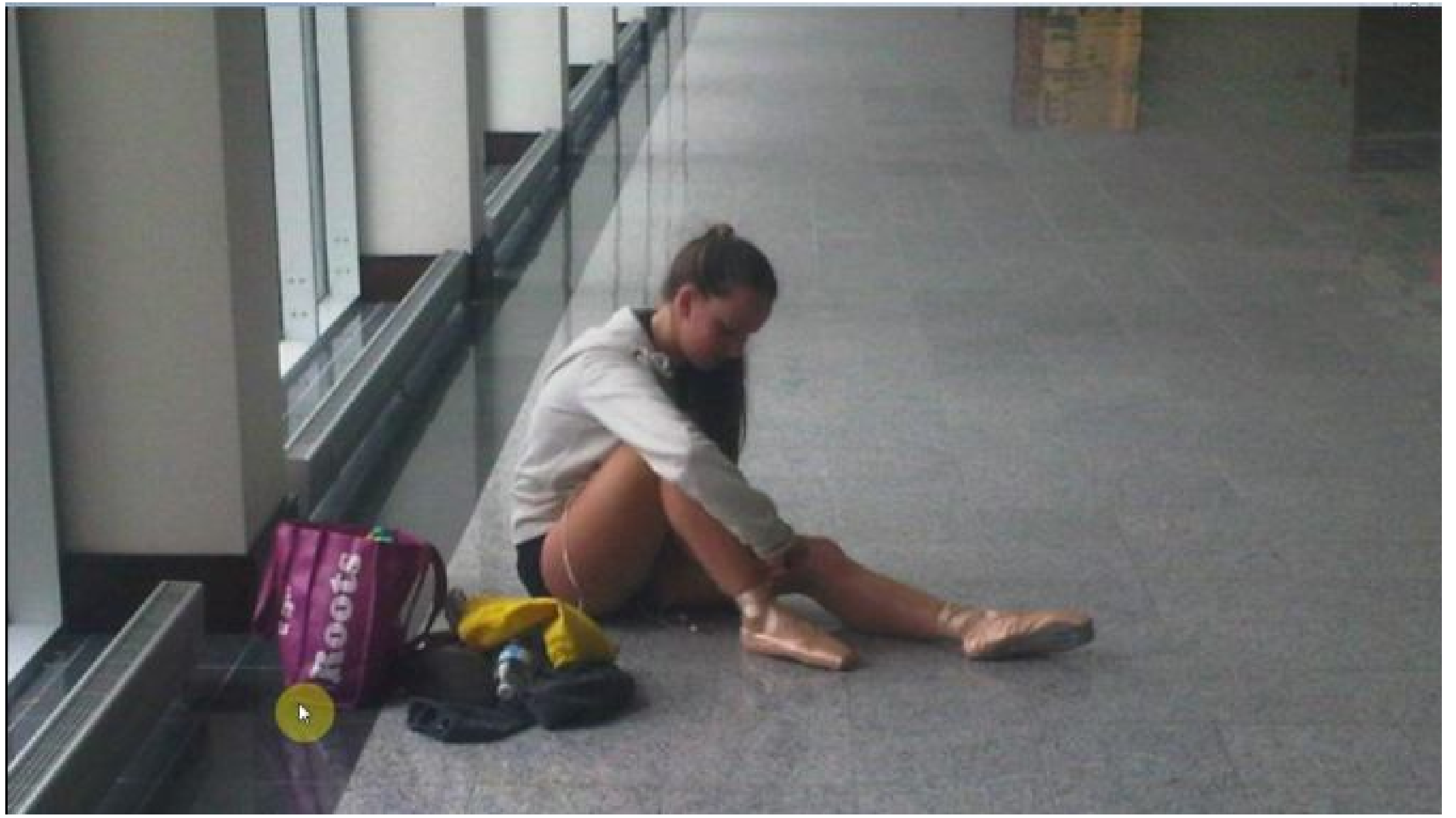
Doctors alarmed that reassessment
medications,

ts harm

KAITLYN ARMSTRONG
Whitby, Ontario

LINDA MORIN
Laval, Quebec

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.

NEW YORK TIMES BESTSELLER

THE BIG FAT SURPRISE

Why Butter, Meat & Cheese Belong in a Healthy Diet

NINA TEICHOLZ

"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."

Disappointing Chinese Vaccine Results Pose Setback for Developing World

Brazil says CoronaVac has an efficacy rate just over 50 percent, much lower than previously announced. More than 380 million doses have already been ordered.

Inspecting vials containing the CoronaVac vaccine, made by the Chinese company Sinovac, at the Butantan Institute in São Paulo, Brazil. Amanda Perobelli/Reuters

Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine

Fernando P. Polack, M.D., Stephen J. Thomas, M.D., Nicholas Kitchin, M.D., Judith Absalon, M.D.,
Alejandra Gurtman, M.D., Stephen Lockhart, D.M., John L. Perez, M.D., Gonzalo Pérez Marc, M.D.,
Edson D. Moreira, M.D., Cristiano Zerbini, M.D., Ruth Bailey, B.Sc., Kena A. Swanson, Ph.D.,
Satrajit Roychoudhury, Ph.D., Kenneth Koury, Ph.D., Ping Li, Ph.D., Warren V. Kalina, Ph.D., David Cooper, Ph.D.,
Robert W. Frenck, Jr., M.D., Laura L. Hammitt, M.D., Özlem Türeci, M.D., Haylene Nell, M.D., Axel Schaefer, M.D.,
Serhat Ünal, M.D., Dina B. Tresnan, D.V.M., Ph.D., Susan Mather, M.D., Philip R. Dormitzer, M.D., Ph.D.,
Uğur Şahin, M.D., Kathrin U. Jansen, Ph.D., and William C. Gruber, M.D., for the C4591001 Clinical Trial Group*

BACKGROUND

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection and the resulting coronavirus disease 2019 (Covid-19) have afflicted tens of millions of people in a worldwide pandemic. Safe and effective vaccines are needed urgently.

METHODS

In an ongoing multinational, placebo-controlled, observer-blinded, pivotal efficacy trial, we randomly assigned persons 16 years of age or older in a 1:1 ratio to receive two doses, 21 days apart, of either placebo or the BNT162b2 vaccine candidate (30 μg per dose). BNT162b2 is a lipid nanoparticle–formulated, nucleoside-modified RNA vaccine that encodes a prefusion stabilized, membrane-anchored SARS-CoV-2 full-length spike protein. The primary end points were efficacy of the vaccine against laboratory-confirmed Covid-19 and safety.

RESULTS

A total of 43,548 participants underwent randomization, of whom 43,448 received injections: 21,720 with BNT162b2 and 21,728 with placebo. There were 8 cases of Covid-19 with onset at least 7 days after the second dose among participants assigned to receive BNT162b2 and 162 cases among those assigned to placebo; BNT162b2 was 95% effective in preventing Covid-19 (95% credible interval, 90.3 to 97.6). Similar vaccine efficacy (generally 90 to 100%) was observed across subgroups defined by age, sex, race, ethnicity, baseline body-mass index, and the presence of coexisting conditions. Among 10 cases of severe Covid-19 with onset after the first dose, 9 occurred in placebo recipients and 1 in a BNT162b2 recipient. The safety profile of BNT162b2 was characterized by short-term, mild-to-moderate pain at the injection site, fatigue, and headache. The incidence of serious adverse events was low and was similar in the vaccine and placebo groups.

CONCLUSIONS

A two-dose regimen of BNT162b2 conferred 95% protection against Covid-19 in persons 16 years of age or older. Safety over a median of 2 months was similar to that of other viral vaccines. (Funded by BioNTech and Pfizer; ClinicalTrials.gov number, NCT04368728.)
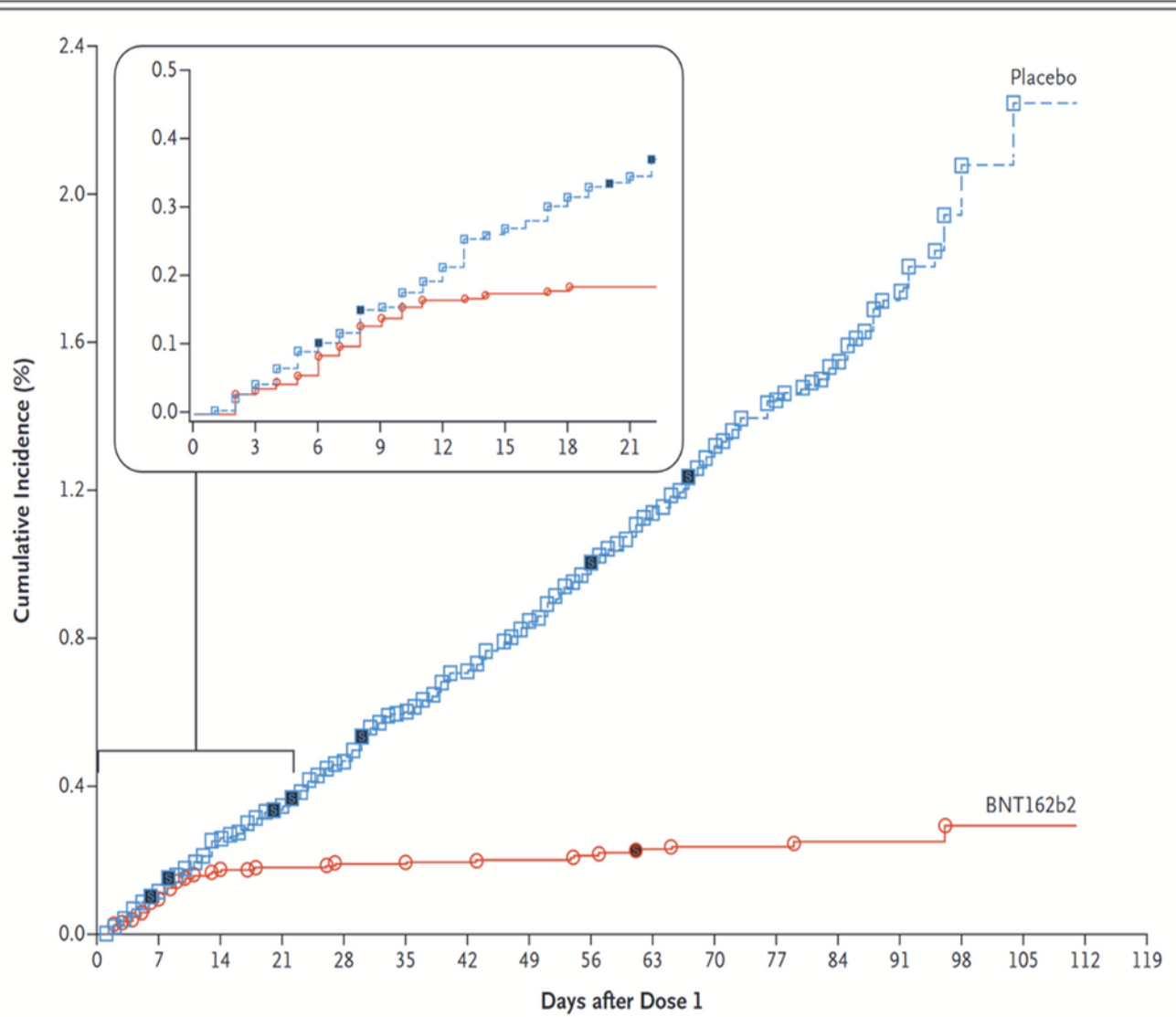
Efficacy End-Point Subgroup	BNT162b2, 30 μg (N=21,669)		Placebo (N=21,686)		VE (95% CI)
	No. of participants	Surveillance time	No. of participants	Surveillance time	
		person-yr (no. at risk)		*person-yr (no. at risk)*	*percent*
Covid-19 occurrence					
After dose 1	50	4.015 (21,314)	275	3.982 (21,258)	82.0 (75.6–86.9)
After dose 1 to before dose 2	39		82		52.4 (29.5–68.4)
Dose 2 to 7 days after dose 2	2		21		90.5 (61.0–98.9)
≥7 Days after dose 2	9		172		94.8 (89.8–97.6)

Efficacy End-Point Subgroup	BNT162b2 (N=18,198)		Placebo (N=18,325)		Vaccine Efficacy, % (95% CI)†
	No. of Cases	Surveillance Time (No. at Risk)*	No. of Cases	Surveillance Time (No. at Risk)*	
Overall	8	2.214 (17,411)	162	2.222 (17,511)	95.0 (90.0–97.9)
Age group					
16 to 55 yr	5	1.234 (9,897)	114	1.239 (9,955)	95.6 (89.4–98.6)
>55 yr	3	0.980 (7,500)	48	0.983 (7,543)	93.7 (80.6–98.8)
≥65 yr	1	0.508 (3,848)	19	0.511 (3,880)	94.7 (66.7–99.9)
≥75 yr	0	0.102 (774)	5	0.106 (785)	100.0 (–13.1–100.0)
Sex					
Male	3	1.124 (8,875)	81	1.108 (8,762)	96.4 (88.9–99.3)
Female	5	1.090 (8,536)	81	1.114 (8,749)	93.7 (84.7–98.0)
Race or ethnic group‡					
White	7	1.889 (14,504)	146	1.903 (14,670)	95.2 (89.8–98.1)
Black or African American	0	0.165 (1,502)	7	0.164 (1,486)	100.0 (31.2–100.0)
All others	1	0.160 (1,405)	9	0.155 (1,355)	89.3 (22.6–99.8)
Hispanic or Latinx	3	0.605 (4,764)	53	0.600 (4,746)	94.4 (82.7–98.9)
Non-Hispanic, non-Latinx	5	1.596 (12,548)	109	1.608 (12,661)	95.4 (88.9–98.5)
Country					
Argentina	1	0.351 (2,545)	35	0.346 (2,521)	97.2 (83.3–99.9)
Brazil	1	0.119 (1,129)	8	0.117 (1,121)	87.7 (8.1–99.7)
United States	6	1.732 (13,359)	119	1.747 (13,506)	94.9 (88.6–98.2)

* Surveillance time is the total time in 1000 person-years for the given end point across all participants within each group at risk for the end point. The time period for Covid-19 case accrual is from 7 days after the second dose to the end of the surveillance period.

‡ The confidence interval (CI) for vaccine efficacy is derived according to the Clopper–Pearson method, adjusted for surveillance time.

Best selling stats
book of all times
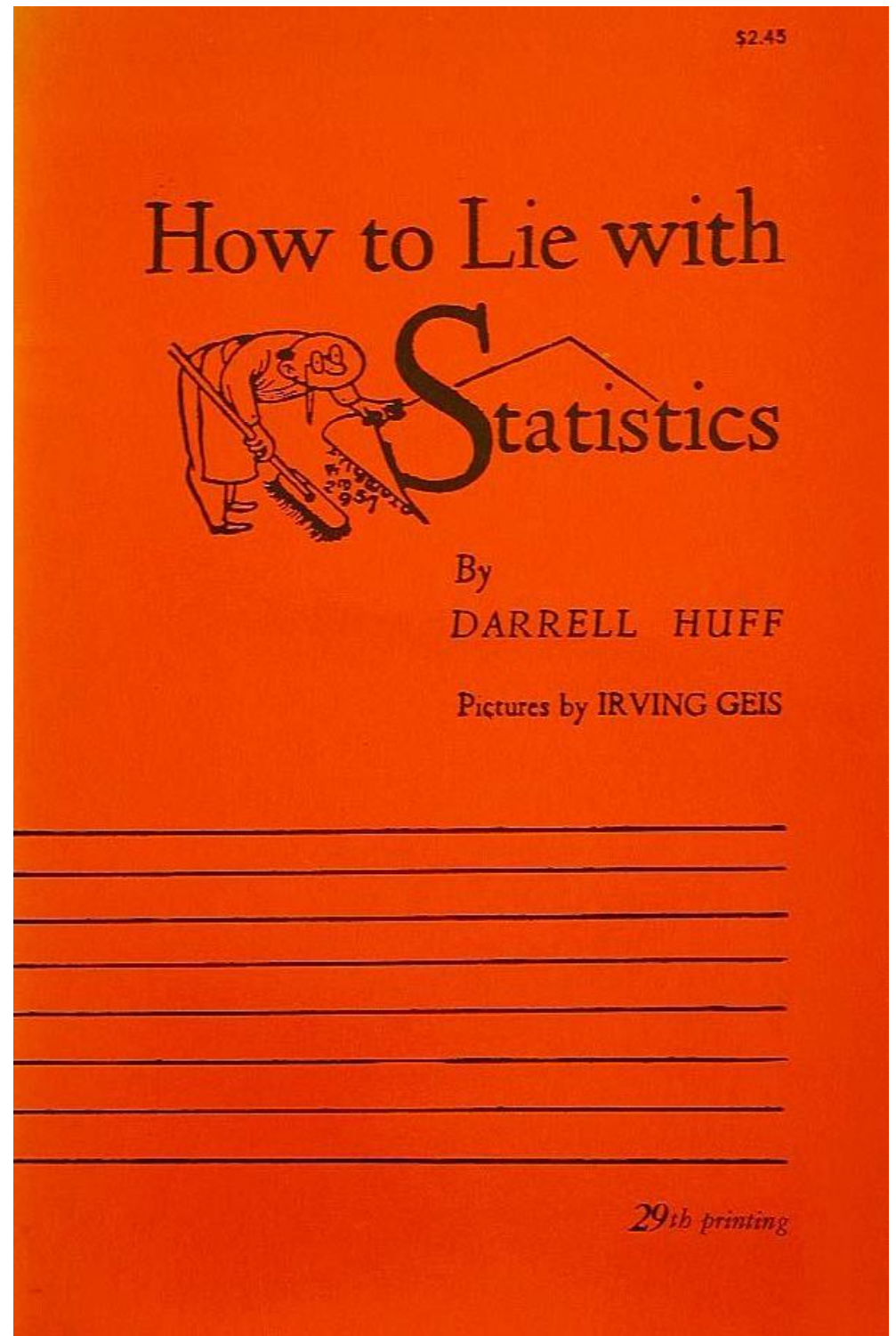
MORE **DAMNED LIES AND STATISTICS**

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

Bad Science
Ben Goldacre
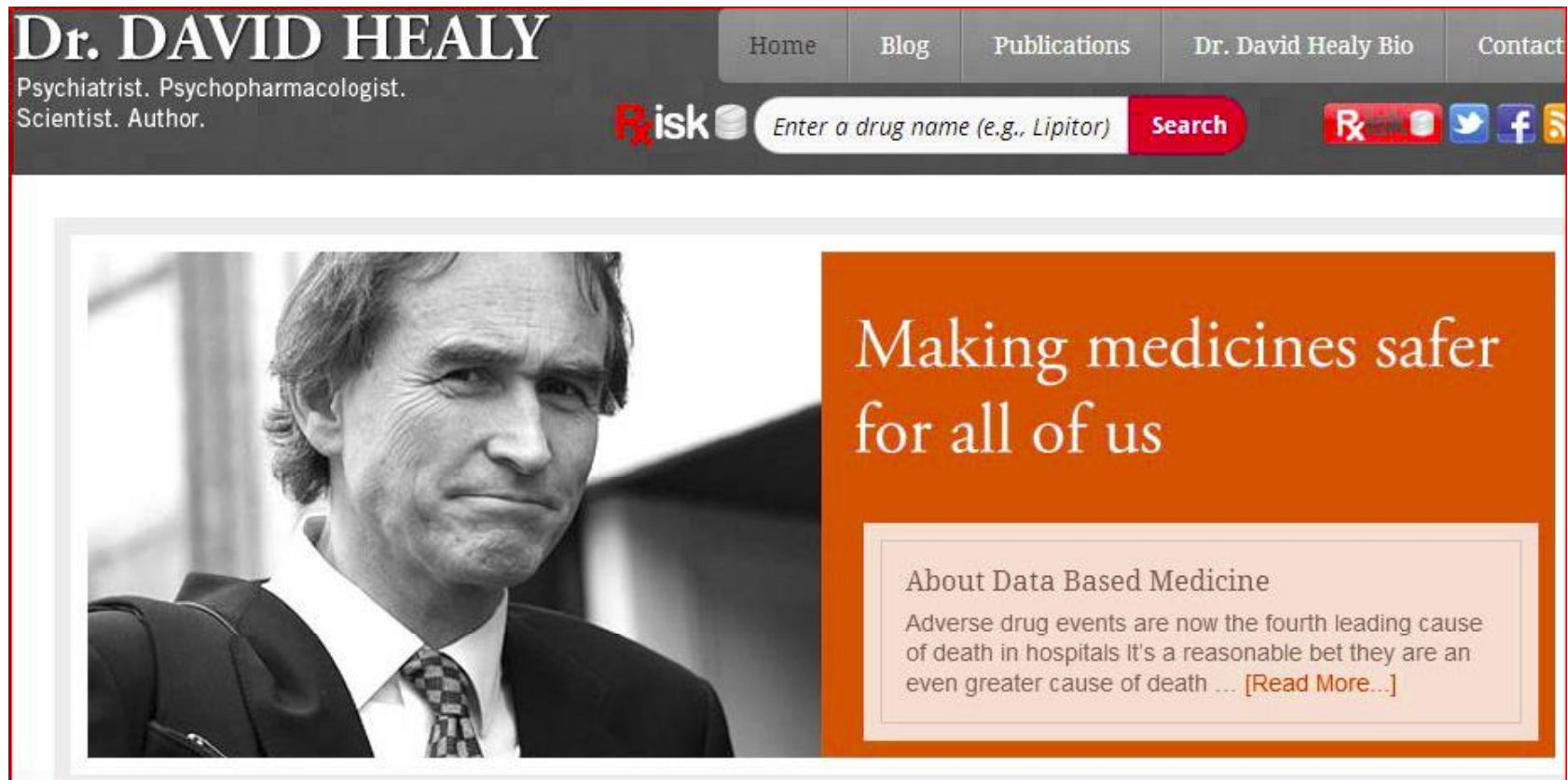
'A fine lesson
in how to
skewer the
enemies of
reason and
the peddlers
of cant and
half-truths'
The Economist

✓ 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):

PHARMA GEDDON
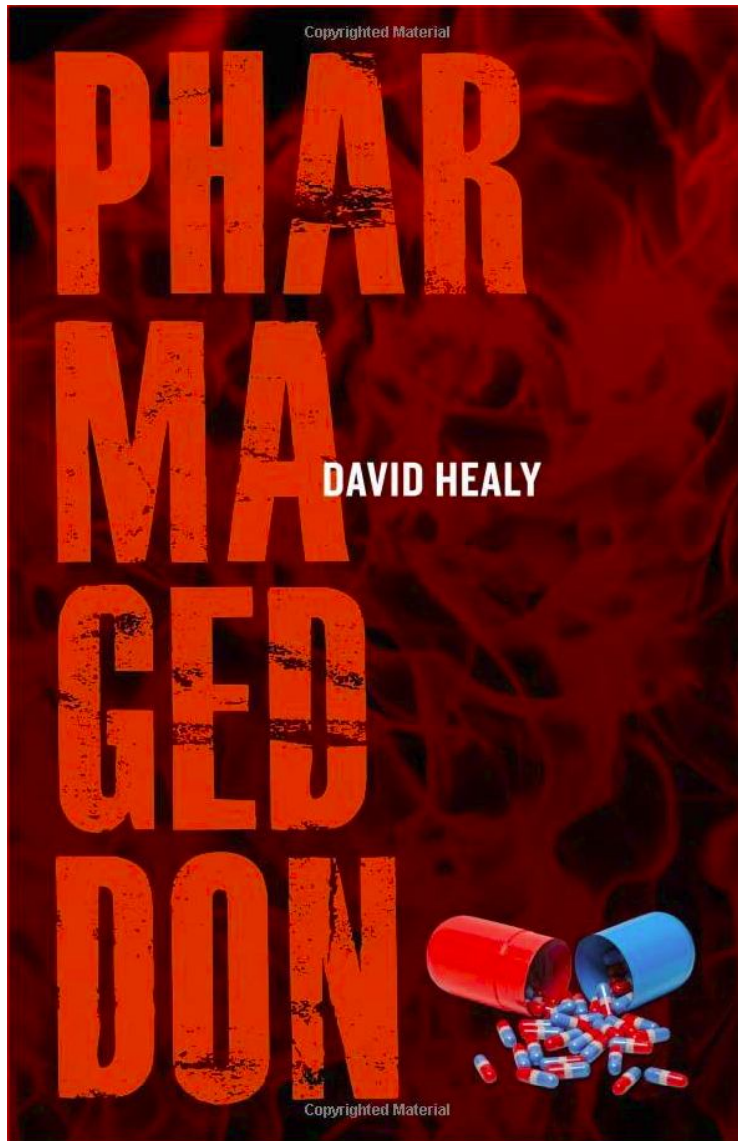
DAVID HEALY

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

JUDEA PEARL

WINNER OF THE TURING AWARD

AND DANA MACKENZIE

THE BOOK OF WHY

α ——————————————➤ β

THE NEW SCIENCE OF CAUSE AND EFFECT

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

thestar.com

60 EARTH HOUR

⊙ thestar.c
Advanced S

HOME   NEWS   OPINION   BUSINESS   SPORTS   ENTERTAINMENT   LIVING   TRAVEL   WHEELS   H

Toronto & GTA | Ontario | Canada | World | Comics | Contests | Crosswords | Horoscopes | Lo

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.

$$Pr[\tau_a < 1, \tau_b < 1] = \Phi_2(\Phi^{-}(F_a(1)), \Phi^{-}(F_b(1)), \gamma)$$

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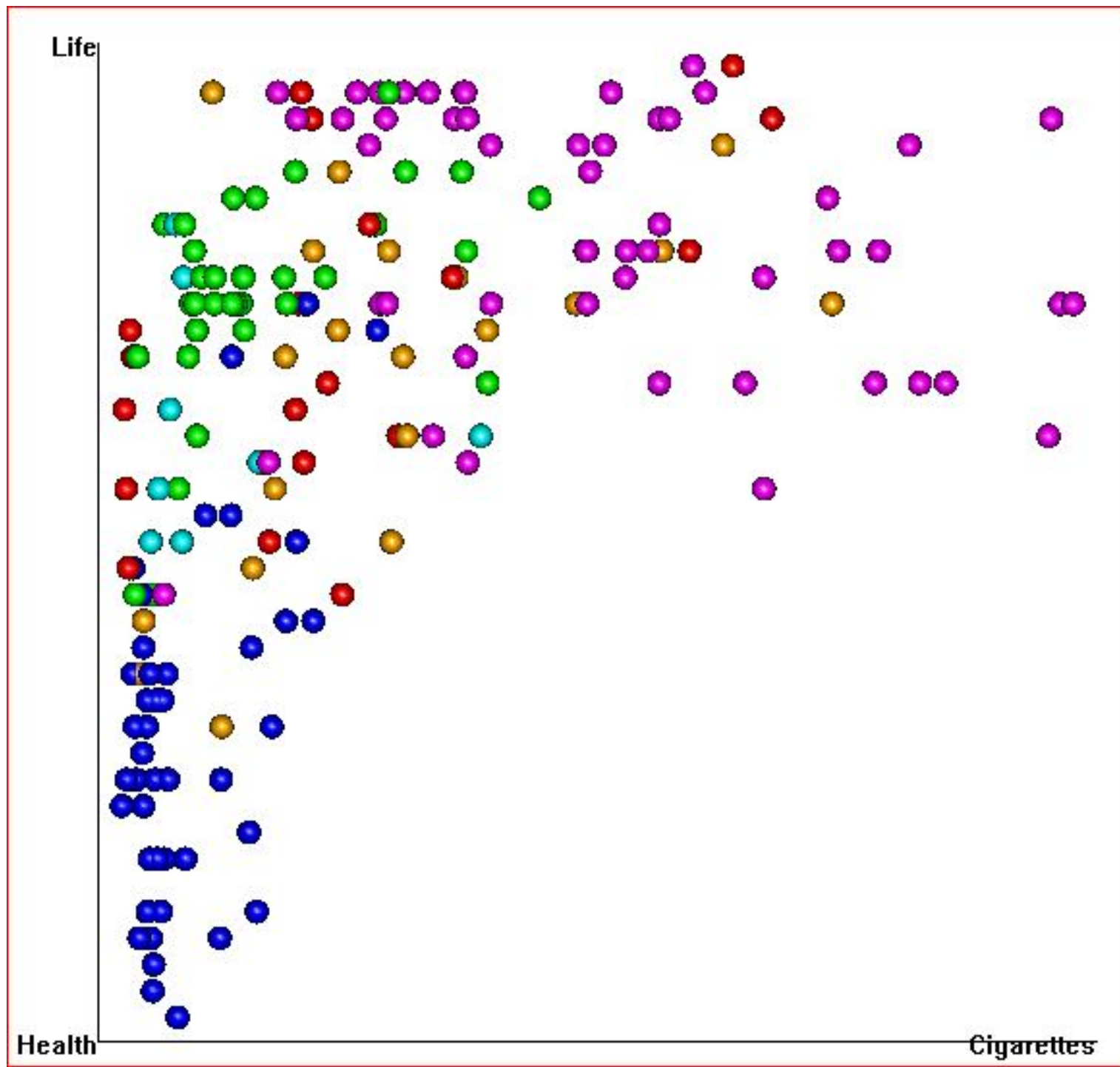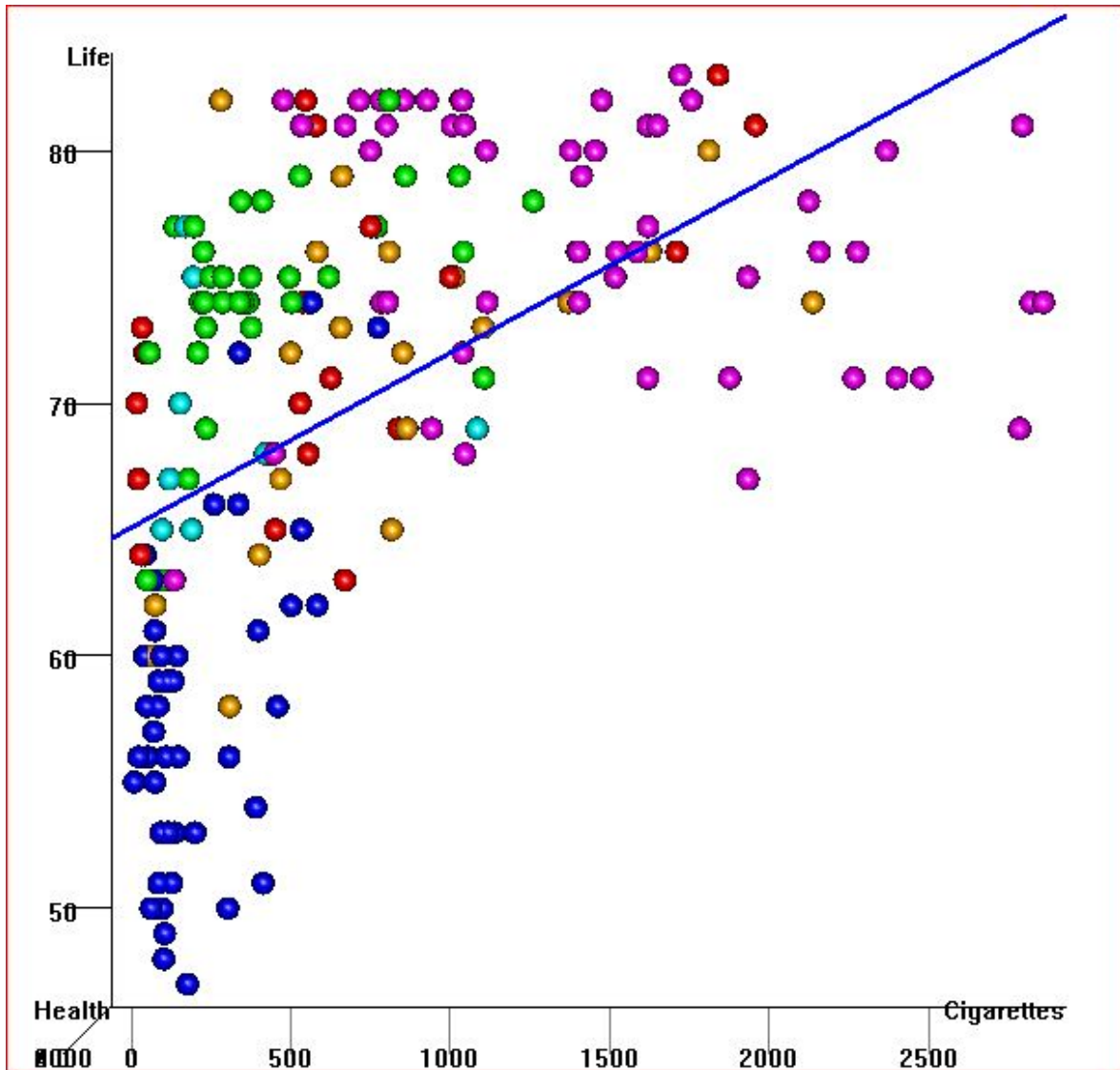
Coefficients	Estimate	Std.Error	DF	t-value	p-value
(Intercept)	65.075840	0.855974	183	76.025515	<.00001
Cigarettes	0.006915	0.000855	183	8.090493	<.00001

```
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
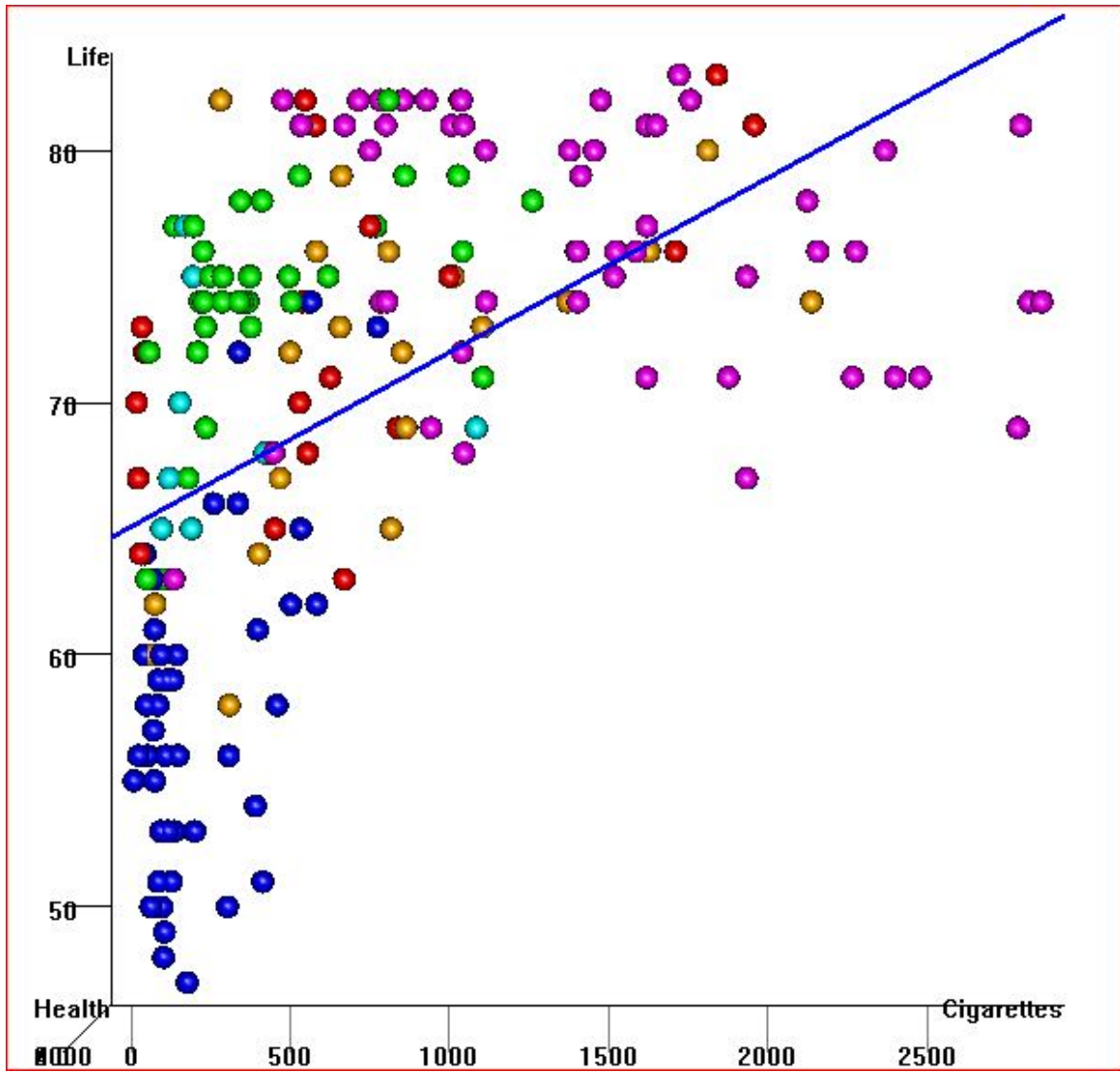
A good statistician would tell you that this is ridiculous.

There's obvious curvature in the relationship
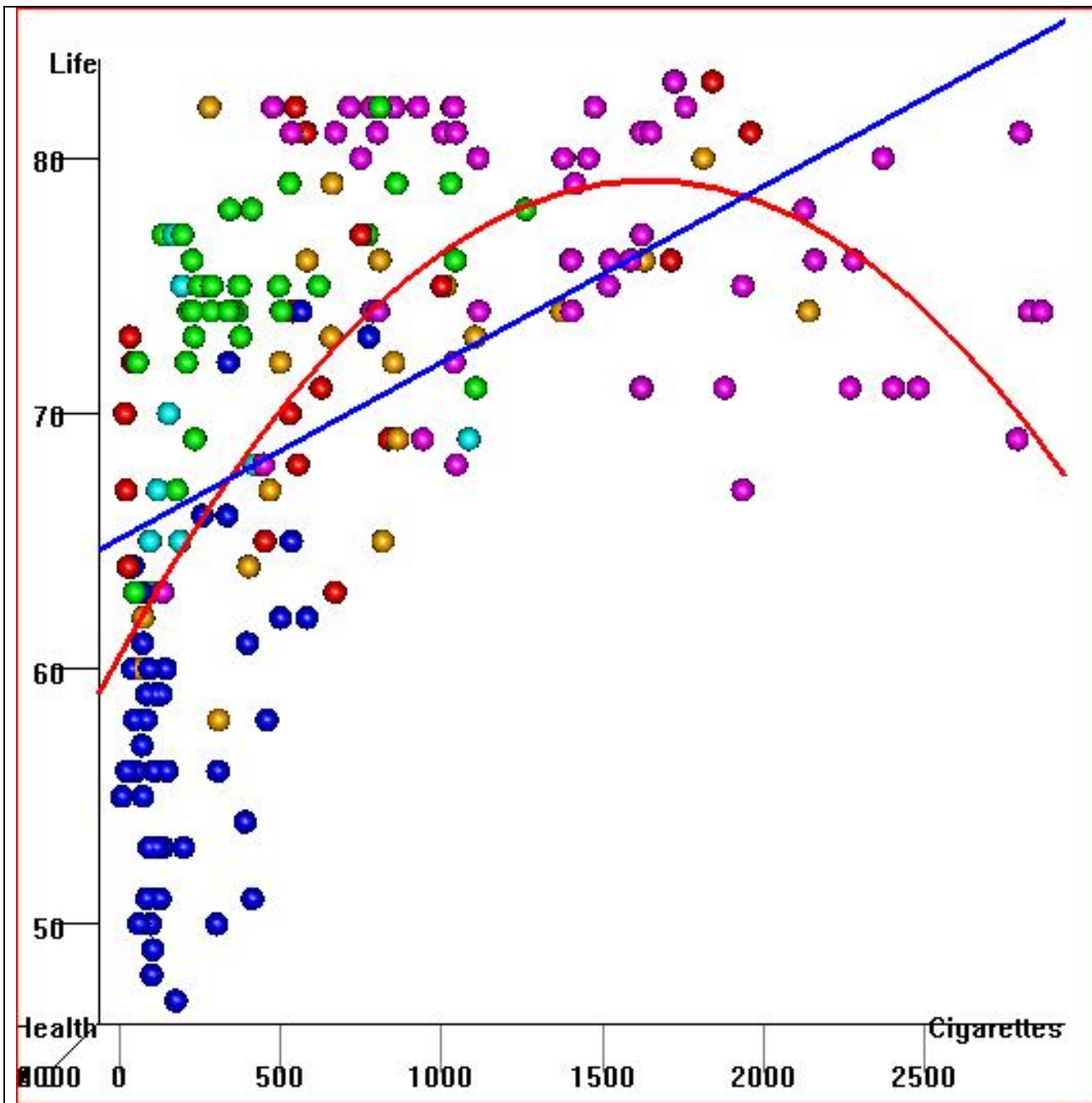
Fitting a quadratic model and maximizing the quadratic shows that

4.495 cigarettes/day

is actually optimal

What's the problem?
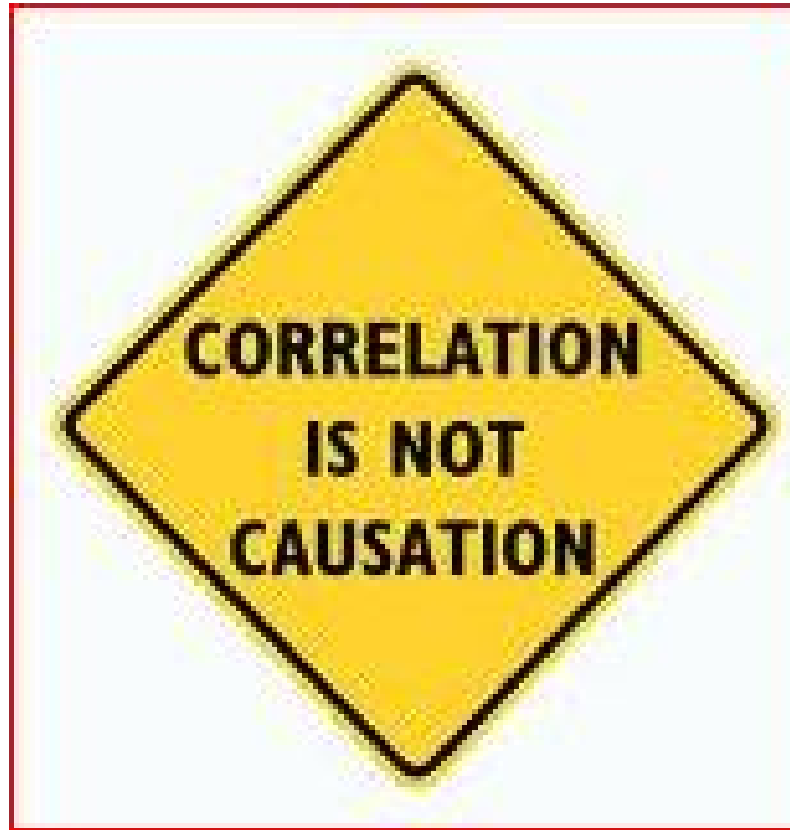
Maybe it isn't smoking that's responsible for higher life expectancies.

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.

Qf $X \iff Y$
what can it mean?

1) $X \Rightarrow Y$

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

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

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

Maybe

a) Directly $X \Rightarrow Y$

b) Through mediating factor(s)

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

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)

$$X \underset{\Rightarrow M_2 \nearrow}{\overset{\nearrow M_1 \Rightarrow}{}} Y$$

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)

$$X \overset{\nearrow M_1 \searrow}{\underset{\searrow M_2 \nearrow}{}} Y$$

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)

$$X \underset{\Rightarrow M_2 \nearrow}{\overset{\nearrow M_1 \Rightarrow}{\Rightarrow}} Y$$

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) $Z \underset{\Rightarrow}{\Rightarrow} \begin{array}{c} X \\ Y \end{array}$ Confounding factor(s)

Maybe

a) Directly $X \Rightarrow Y$

b) Through mediating factor(s)
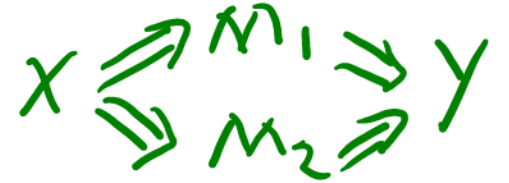
$X \underset{\Rightarrow M_2 \Rightarrow}{\overset{\Rightarrow M_1 \Rightarrow}{}} Y$

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

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

$$X \overset{\nearrow M_1 \searrow}{\underset{\searrow M_2 \nearrow}{}} Y$$

2) $Y \Rightarrow X$

3) $Z \overset{\nearrow X}{\underset{\searrow Y}{}}$ Confounding factor(s)

a) Z known + measurable

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

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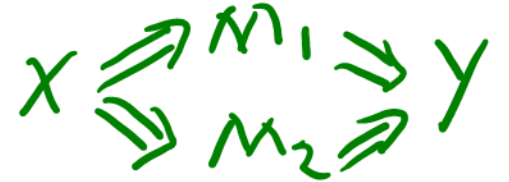a) Directly $X \Rightarrow Y$
b) Through mediating factor(s)

$$X \overset{\nearrow M_1 \searrow}{\underset{\searrow M_2 \nearrow}{}} Y$$

2) $Y \Rightarrow X$

3) $Z \overset{\Rightarrow X}{\underset{\Rightarrow Y}{}}$ Confounding factor(s)

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

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

<u>Maybe</u>
a) Directly $X \Rightarrow Y$
b) Through <u>mediating</u> <u>factor(s)</u>
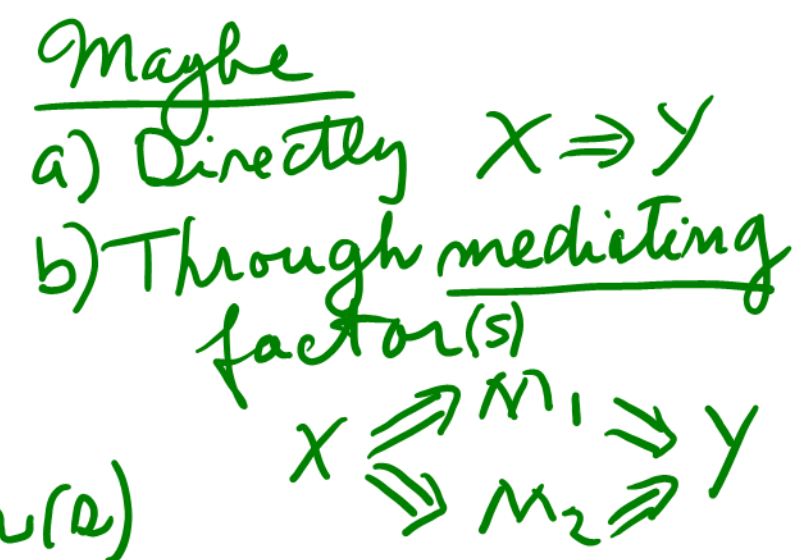
$$X \underset{\Longrightarrow}{\overset{\Longrightarrow}{}} \begin{matrix} M_1 \Rightarrow \\ M_2 \end{matrix} Y$$

2) $Y \Rightarrow X$

3) $Z \underset{\Longrightarrow}{\overset{\Longrightarrow}{}} \begin{matrix} X \\ Y \end{matrix}$ Confounding factor(s)

a) Z known + measurable
b) Z " but hard to measure

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

$\underline{\text{Maybe}}$
a) Directly $X \Rightarrow Y$
b) Through $\underline{\text{mediating}}$ factor(s)

$$X \begin{array}{c} \nearrow M_1 \searrow \\ \searrow M_2 \nearrow \end{array} Y$$

2) $Y \Rightarrow X$

3) $Z \begin{array}{c} \nearrow X \\ \searrow Y \end{array}$ Confounding factor(s)

a) Z known + measurable
b) Z " but hard to measure $\longrightarrow$ adjust for measurement error – SEMs

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

<u>Maybe</u>

a) Directly $X \Rightarrow Y$

b) Through <u>mediating</u> factor(s)

$$X \begin{array}{c} \nearrow M_1 \searrow \\ \searrow M_2 \nearrow \end{array} Y$$

2) $Y \Rightarrow X$

3) $Z \begin{array}{c} \nearrow X \\ \searrow Y \end{array}$ Confounding factor(s)

a) Z known & measurable

b) Z " but hard to measure

c) Z unknown

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

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

$X \genfrac{}{}{0pt}{}{\nearrow M_1 \searrow}{\searrow M_2 \nearrow} Y$

2) $Y \Rightarrow X$

3) $Z \genfrac{}{}{0pt}{}{\nearrow X}{\searrow Y}$ 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

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

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

$$X \overset{\nearrow M_1 \searrow}{\underset{\searrow M_2 \nearrow}{}} Y$$

2) $Y \Rightarrow X$

3) $Z \overset{\Rightarrow X}{\underset{\Rightarrow Y}{}}$ 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 $\rightarrow$ use longitudinal or nested data

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

Maybe
a) Directly $X \Rightarrow Y$
b) Through mediating factor(s)
$$X \underset{\Rightarrow}{\overset{\nearrow}{\Rightarrow}} \overset{M_1 \Rightarrow}{\underset{M_2 \nearrow}{}} Y$$

2) $Y \Rightarrow X$

3) $Z \overset{\Rightarrow X}{\underset{\Rightarrow Y}{}}$ 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

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

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

$$X \underset{\Longrightarrow M_2}{\overset{\nearrow M_1}{\Longrightarrow}} \searrow Y$$

2) $Y \Rightarrow X$

3) $Z \overset{\Rightarrow X}{\underset{\Longrightarrow Y}{}}$ 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

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

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

$$X \begin{array}{c} \nearrow M_1 \searrow \\ \searrow M_2 \nearrow \end{array} Y$$

2) $Y \Rightarrow X$

3) $Z \begin{array}{c} \nearrow X \\ \searrow Y \end{array}$ 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

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

2) $Y \Rightarrow X$

3) $Z \overset{\Rightarrow X}{\underset{\Rightarrow Y}{}}$ 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)

$X \overset{\nearrow M_1 \Rightarrow}{\underset{\Rightarrow M_2 \nearrow}{}} Y$

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

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

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

$$X \begin{smallmatrix} \nearrow M_1 \searrow \\ \searrow M_2 \nearrow \end{smallmatrix} Y$$

2) $Y \Rightarrow X$

3) $Z \begin{smallmatrix} \nearrow X \\ \searrow Y \end{smallmatrix}$ Confounding factor(s)

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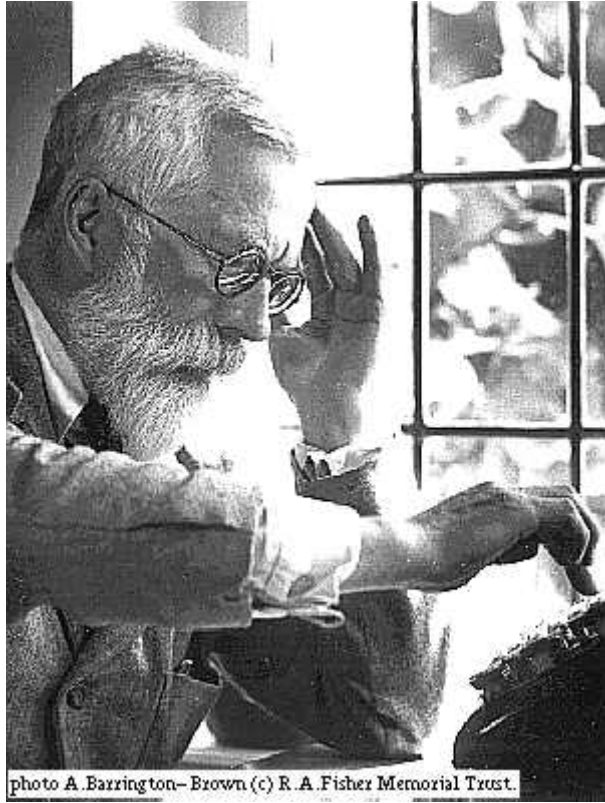c) Z unknown
d) There are clusters in which Z is constant

4) Chance

5) Selection

- 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):

photo A.Barrington–Brown (c) R.A.Fisher Memorial Trust.

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*

"A calculated risk": the Salk polio vaccine field trials of 1954

Marcia Meldrum, DeWitt Stetten memorial fellow in the history of the biomedical sciences

‣ Author information ‣ Article notes ‣ Copyright and License information Disclaimer

The polio vaccine field trials of 1954, sponsored by the National Foundation for Infantile Paralysis (March of Dimes), are among the largest and most publicised clinical trials ever undertaken. Across the United States, 623 972 schoolchildren were injected with vaccine or placebo, and more than a million others participated as "observed" controls. The results, announced in 1955, showed good statistical evidence that Jonas Salk's killed virus preparation was 80-90% effective in preventing paralytic poliomyelitis.[1]

The statistical design used in this great experiment was singular, prompting criticism at the time and since. Eighty four test areas in 11 states used the textbook model: in a randomised, blinded design all participating children in the first three grades of school (ages 6-9) received injections of either vaccine or placebo and were observed for evidence of the disease. But 127 test areas in 33 states used an "observed control" design: participating children in the second grade (ages 7-8) received injections of vaccine; no placebo was given, and children in all three grades were then observed for the duration of the polio "season."[1]

The use of the dual protocol illustrates both the power and the limitations of the randomised clinical trial to legitimate therapeutic claims. The placebo controlled trials were necessary to define the Salk vaccine—introduced by a lay organisation that has taken an activist position against the counsel of its virological advisers—as the product of scientific medicine. The observed control trials were essential to maintaining public support for the vaccine as the product of lay faith and investment in science. Here I examine the process by which the trial design was negotiated and the roles of the several actors.

Disappointing Chinese Vaccine Results Pose Setback for Developing World

Brazil says CoronaVac has an efficacy rate just over 50 percent, much lower than previously announced. More than 380 million doses have already been ordered.

By Sui-Lee Wee and Ernesto Londoño

Jan. 13, 2021 Updated 8:29 a.m. ET

Scientists in Brazil have downgraded the efficacy of a Chinese coronavirus vaccine that they hailed as a major triumph last week, diminishing hopes for a shot that could be quickly produced and easily distributed to help the developing world.

Officials at the Butantan Institute in São Paulo said on Tuesday that a trial conducted in Brazil showed that the CoronaVac vaccine, manufactured by the Beijing-based company Sinovac, had an efficacy rate just over 50 percent. That rate, slightly above the benchmark that the World Health Organization has said would make a vaccine effective for general use, was far below the 78 percent level announced last week.

The implications could be significant for a vaccine that is crucial to China's global health diplomacy. At least 10 countries have ordered more than 380 million doses of the Sinovac inoculation, CoronaVac, though regulatory agencies have yet to fully approve it.

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

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

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

$$X \overset{\nearrow M_1 \searrow}{\underset{\searrow M_2 \nearrow}{}} Y$$

2) $Y \Rightarrow X$

3) $Z \overset{\Rightarrow X}{\underset{\Rightarrow Y}{}}$ 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

- 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) $Z \overset{\Rightarrow X}{\underset{\Rightarrow Y}{}}$

    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) $Z \begin{array}{c} \nearrow X \\ \searrow Y \end{array}$

    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) $Z \begin{cases} \Rightarrow X \\ \Rightarrow Y \end{cases}$

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) $Z \begin{array}{c} \nearrow X \\ \searrow Y \end{array}$

    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) $Z \begin{cases} \Rightarrow X \\ \Rightarrow Y \end{cases}$

   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) $Z \begin{array}{c} \nearrow X \\ \searrow Y \end{array}$

    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) $Z \begin{cases} \Rightarrow X \\ \Rightarrow Y \end{cases}$

   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) $Z \underset{\Rightarrow Y}{\overset{\Rightarrow X}{}}$

    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) $Z \underset{\searrow Y}{\overset{\nearrow X}{}}$

   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) ?

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) $Z \nearrow X$
   $Z \searrow Y$

   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) $Z \begin{array}{c} \nearrow X \\ \searrow Y \end{array}$

   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
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) $Z \begin{cases} \Rightarrow X \\ \Rightarrow Y \end{cases}$

   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)

$\left.\begin{array}{c} \\ \\ \\ \end{array}\right\}$ 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:

Questions	**Causal** what would happen if ...?		
	Predictive passive guessing		

The fundamental
2 x 2 table of statistics

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:
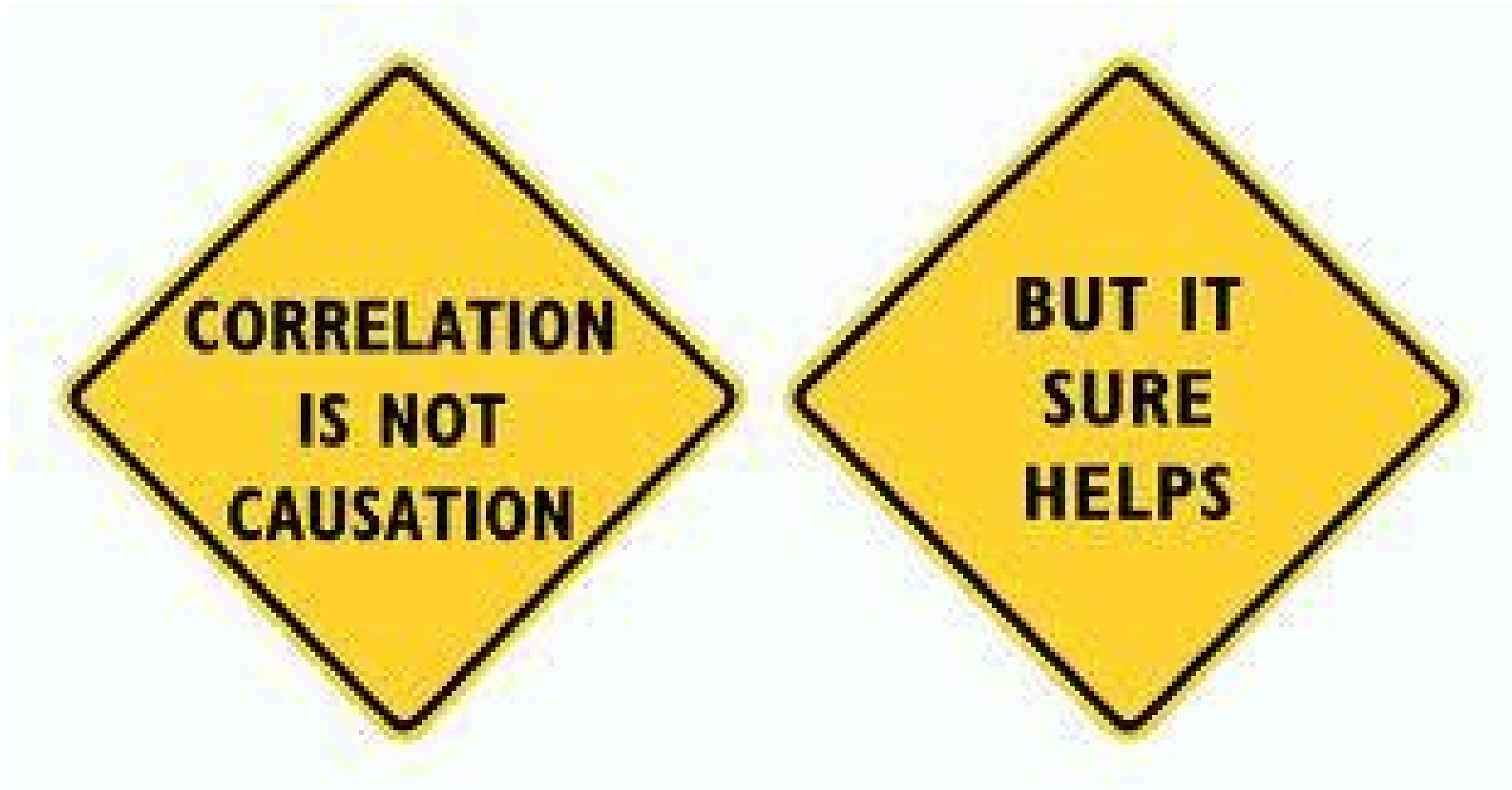
How should we react to them?

(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.