Thinking About Data

MMED
African Institute for the Mathematical Sciences
Muizenberg, South Africa
May, 2019

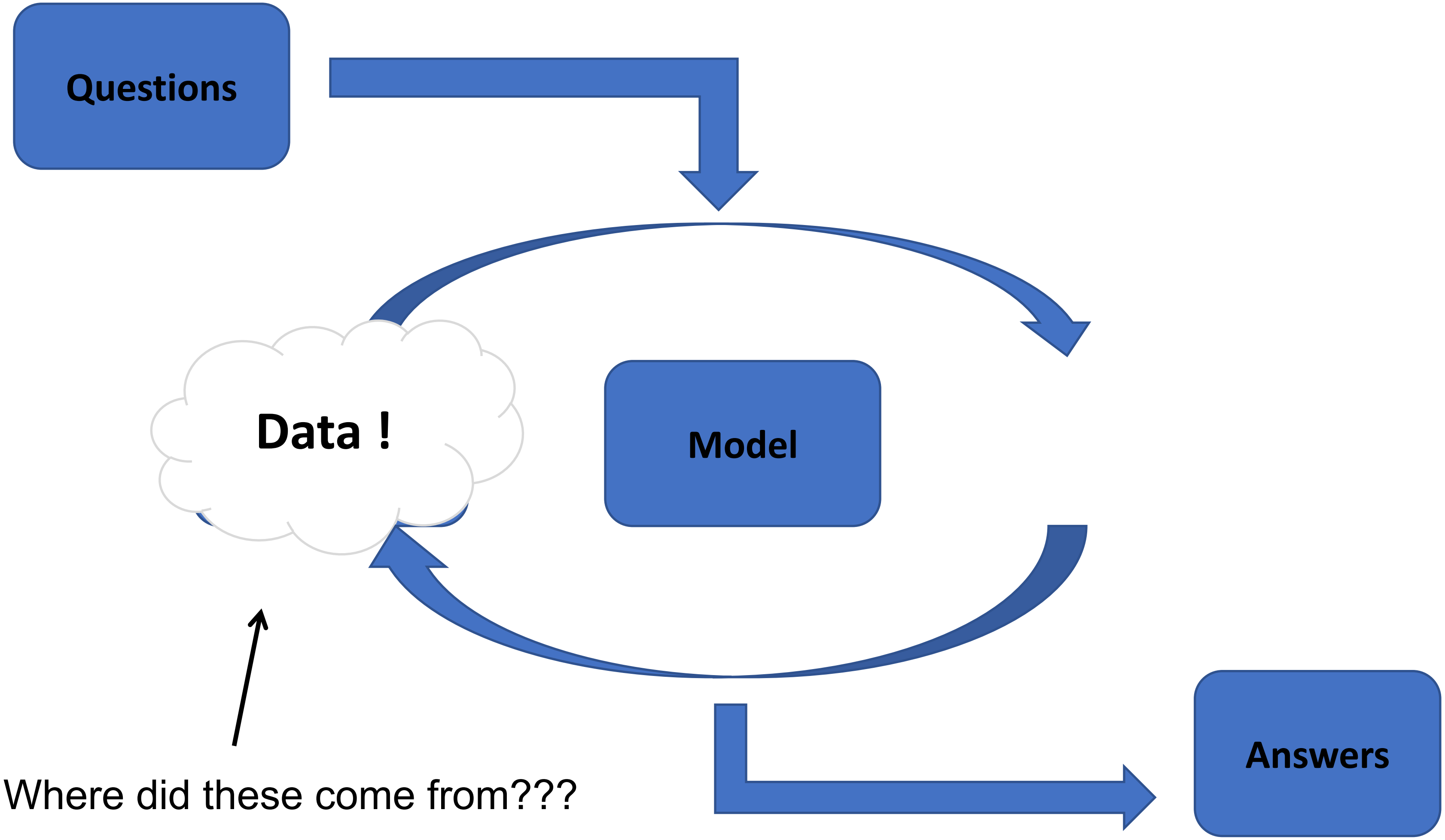
Jim Scott, Ph.D, M.A., M.P.H.

Slide set citation: https://doi.org/10.6084/m9.figshare.5044615.v4

Goals:

- Recognize that data are better than anecdotes
- Understand that how the data are collected matters
- Be aware of variability – don’t be fooled by it
- Consider confounding and other forms of bias
- Healthy skepticism

Why do we care about data?



Where did these come from???

“Data, data, data!”, he cried impatiently. “I can’t make bricks without clay”

- Sherlock Holmes

Source: Statistics 3rd, ed. Pisani Purves, Freedman

The New York Times

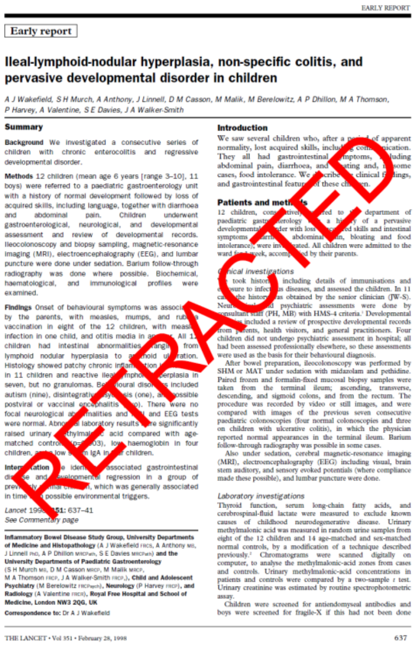
February 3, 2010

Journal Retracts 1998 Paper Linking Autism to Vaccines

By GARDINER HARRIS

A prominent British medical journal on Tuesday retracted a 1998 research paper that set off a sharp decline in vaccinations in Britain after the paper's lead author suggested that vaccines could cause autism.

The retraction by The Lancet is part of a reassessment that has lasted for years of the scientific methods and financial conflicts of Dr. Andrew Wakefield, who contended that his research showed that the combined measles, mumps and rubella vaccine may be unsafe.



EARLY REPORT

Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith

RETRACTED

Summary

Background We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

Methods 12 children (mean age 6 years [range 3–10], 11 boys) were referred to a paediatric gastroenterology unit with a history of normal development followed by loss of acquired skills, including language, together with diarrhoea and abdominal pain. Children underwent gastroenterological, neurological, and developmental assessment and review of developmental records. Ileocolonoscopy and biopsy sampling, magnetic-resonance imaging (MRI), electroencephalography (EEG), and lumbar puncture were done under sedation. Barium follow-through radiography was done where possible. Biochemical, haematological, and immunological profiles were examined.

Findings Onset of behavioural symptoms was associated, by the parents, with measles, mumps, and rubella vaccination in eight of the 12 children, with measles infection in one child, and otitis media in another. All 12 children had intestinal abnormalities, ranging from lymphoid nodular hyperplasia to aphthoid ulceration. Histology showed patchy chronic inflammation in the colon in 11 children and reactive ileal lymphoid hyperplasia in seven, but no granulomas. Behavioural disorders included autism (nine), disintegrative psychosis (one), and possible postviral or vaccinal encephalitis (two). There were no focal neurological abnormalities and MRI and EEG tests were normal. Abnormal laboratory results were significantly raised urinary methylmalonic acid compared with age-matched controls (p=0·003), low haemoglobin in four children, and a low serum IgA in four children.

Interpretation We identified associated gastrointestinal disease and developmental regression in a group of previously normal children, which was generally associated in time with possible environmental triggers.

Lancet 1998; **351:** 637–41

See Commentary page

Inflammatory Bowel Disease Study Group, University Departments of Medicine and Histopathology (A J Wakefield FRCS, A Anthony MB, J Linnell PhD, A P Dhillon MRCPath, S E Davies MRCPath) **and the University Departments of Paediatric Gastroenterology** (S H Murch MB, D M Casson MRCP, M Malik MRCP, M A Thomson FRCP, J A Walker-Smith FRCP,), **Child and Adolescent Psychiatry** (M Berelowitz FRCPsych), **Neurology** (P Harvey FRCP), **and Radiology** (A Valentine FRCR), **Royal Free Hospital and School of Medicine, London NW3 2QG, UK**

Correspondence to: Dr A J Wakefield

Introduction

We saw several children who, after a period of apparent normality, lost acquired skills, including communication. They all had gastrointestinal symptoms, including abdominal pain, diarrhoea, and bloating and, in some cases, food intolerance. We describe the clinical findings, and gastrointestinal features of these children.

Patients and methods

12 children, consecutively referred to the department of paediatric gastroenterology with a history of a pervasive developmental disorder with loss of acquired skills and intestinal symptoms (diarrhoea, abdominal pain, bloating and food intolerance), were investigated. All children were admitted to the ward for 1 week, accompanied by their parents.

Clinical investigations

We took histories, including details of immunisations and exposure to infectious diseases, and assessed the children. In 11 cases the history was obtained by the senior clinician (JW-S). Neurological and psychiatric assessments were done by consultant staff (PH, MB) with HMS-4 criteria.[1] Developmental histories included a review of prospective developmental records from parents, health visitors, and general practitioners. Four children did not undergo psychiatric assessment in hospital; all had been assessed professionally elsewhere, so these assessments were used as the basis for their behavioural diagnosis.

After bowel preparation, ileocolonoscopy was performed by SHM or MAT under sedation with midazolam and pethidine. Paired frozen and formalin-fixed mucosal biopsy samples were taken from the terminal ileum; ascending, transverse, descending, and sigmoid colons, and from the rectum. The procedure was recorded by video or still images, and were compared with images of the previous seven consecutive paediatric colonoscopies (four normal colonoscopies and three on children with ulcerative colitis), in which the physician reported normal appearances in the terminal ileum. Barium follow-through radiography was possible in some cases.

Also under sedation, cerebral magnetic-resonance imaging (MRI), electroencephalography (EEG) including visual, brain stem auditory, and sensory evoked potentials (where compliance made these possible), and lumbar puncture were done.

Laboratory investigations

Thyroid function, serum long-chain fatty acids, and cerebrospinal-fluid lactate were measured to exclude known causes of childhood neurodegenerative disease. Urinary methylmalonic acid was measured in random urine samples from eight of the 12 children and 14 age-matched and sex-matched normal controls, by a modification of a technique described previously.[2] Chromatograms were scanned digitally on computer, to analyse the methylmalonic-acid zones from cases and controls. Urinary methylmalonic-acid concentrations in patients and controls were compared by a two-sample *t* test. Urinary creatinine was estimated by routine spectrophotometric assay.

Children were screened for antiendomyseal antibodies and boys were screened for fragile-X if this had not been done

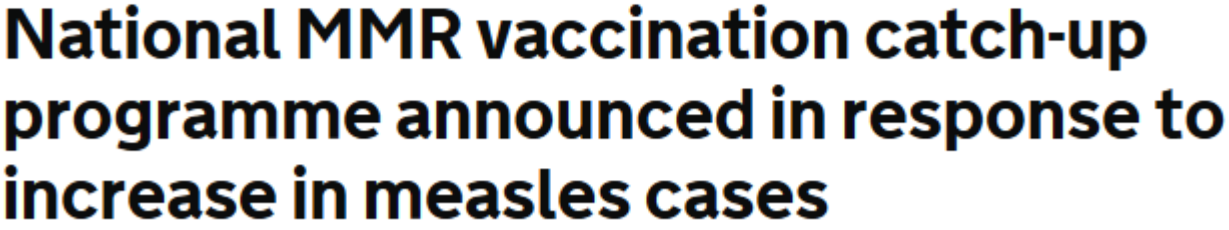
THE LANCET • Vol 351 • February 28, 1998

637

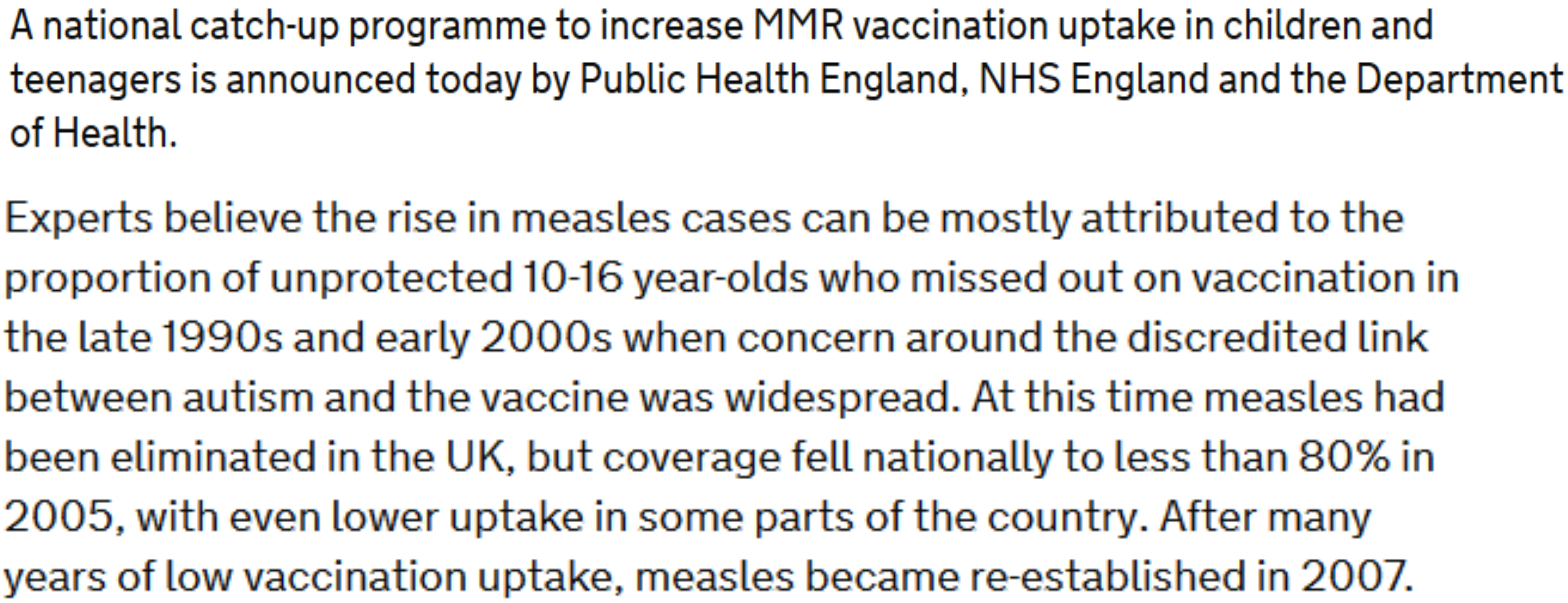
"The story became credible because it was published in The Lancet," Alison Singer, president of the Autism Science Foundation, said Tuesday. "It was in The Lancet, and we really rely on these medical journals."

Consequences

Commons.wikimedia.org



National MMR vaccination catch-up programme announced in response to increase in measles cases



A national catch-up programme to increase MMR vaccination uptake in children and teenagers is announced today by Public Health England, NHS England and the Department of Health.

Experts believe the rise in measles cases can be mostly attributed to the proportion of unprotected 10-16 year-olds who missed out on vaccination in the late 1990s and early 2000s when concern around the discredited link between autism and the vaccine was widespread. At this time measles had been eliminated in the UK, but coverage fell nationally to less than 80% in 2005, with even lower uptake in some parts of the country. After many years of low vaccination uptake, measles became re-established in 2007.

https://www.gov.uk/government/news/national-mmr-vaccination-catch-up-programme-announced-in-response-to-increase-in-measles-cases

How the Data are Collected Matters

- “Always do right. This will gratify some people, and astonish the rest”

  - Mark Twain

- Beware: All data are not created equal

Source: Statistics 3rd, ed. Pisani, Purves, Freedman

1936 Presidential Election

1936 Presidential Election

- 1936 Literary Digest Poll
- Literary Digest had predicted the winner of every US presidential election since 1916.
- In 1936, Literary Digest mailed questionnaires to 10 million people (25% of voters).
- 2.4 million people responded

- Returned questionnaires:

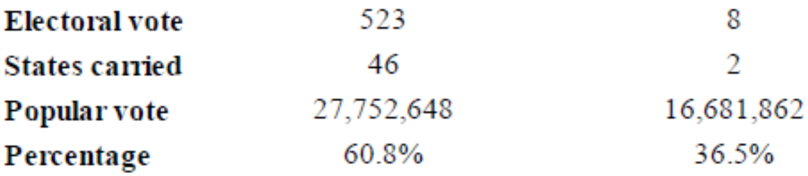
Landon:	**1,293,668**	**57%**
FDR:	**972,897**	**43%**

Source: http://historymatters.gmu.edu/d/5168/

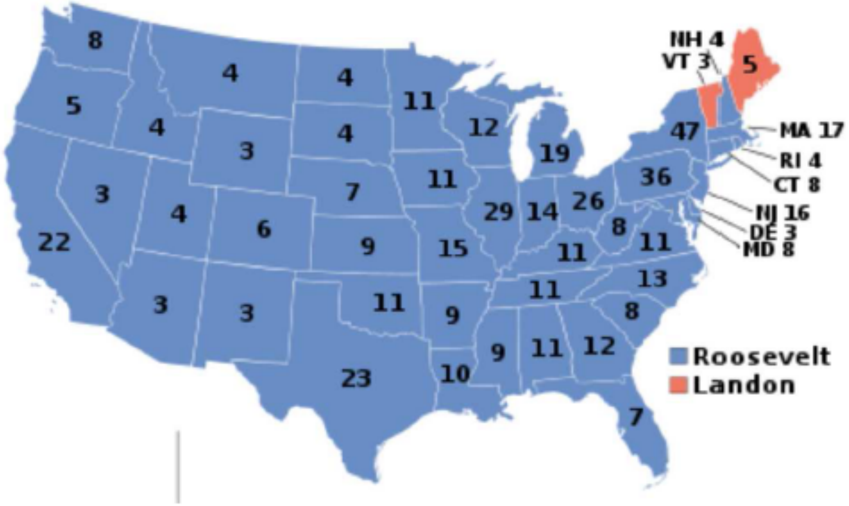
Results

- **Actual Result: Roosevelt 61%, Landon 37%.**

- **One of the biggest landslides in U.S. history**



Electoral vote	523	8
States carried	46	2
Popular vote	27,752,648	16,681,862
Percentage	60.8%	36.5%



What went wrong?

- How were the data collected?

  - Those who received the questionnaire were systematically different than those who didn’t
    - 10 million sent out (~25% of voters)
    - 2.3 million returned – sample of convenience
      - Not representative
  - Sampling frame:
    - Telephone books
    - Automobile registries

Bias and Variability

- Sample size doesn't matter if the data collection scheme is flawed

Observed value = The TRUTH + **Bias** + *Chance Error*

Another example?

- IBM created Watson
  - Uses data to make predictions – finds patterns in knowledge database
- Trained Watson to play Jeopardy
- Crushed all humans
- Triumph for A.I.
  - …..mostly…..

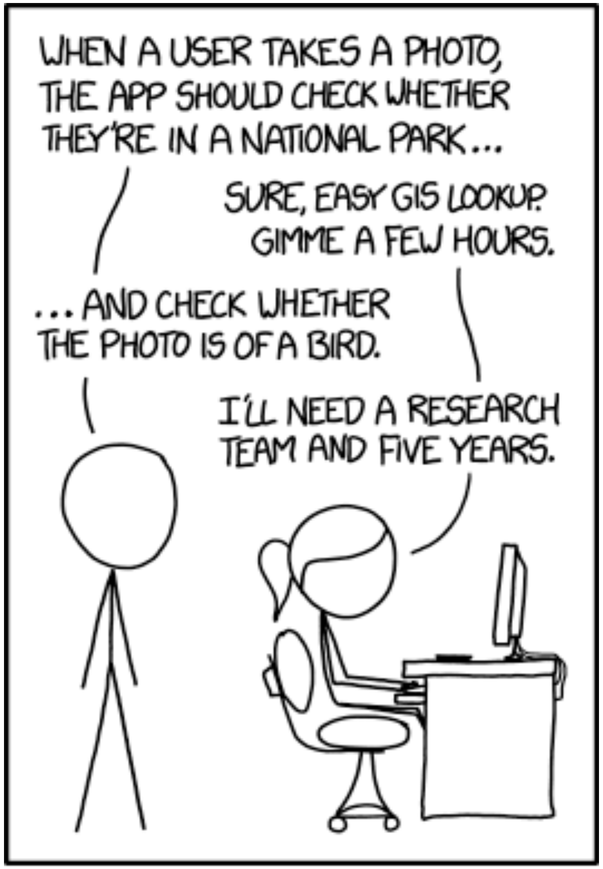
U.S. Cities

Its largest airport is named for a World War II hero; its second largest, for a World War II battle

Watson:
What is…… Toronto(?!)

Oops

Big Data Limitations


WHEN A USER TAKES A PHOTO, THE APP SHOULD CHECK WHETHER THEY'RE IN A NATIONAL PARK...
SURE, EASY GIS LOOKUP. GIMME A FEW HOURS.
...AND CHECK WHETHER THE PHOTO IS OF A BIRD.
I'LL NEED A RESEARCH TEAM AND FIVE YEARS.

IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

Variability

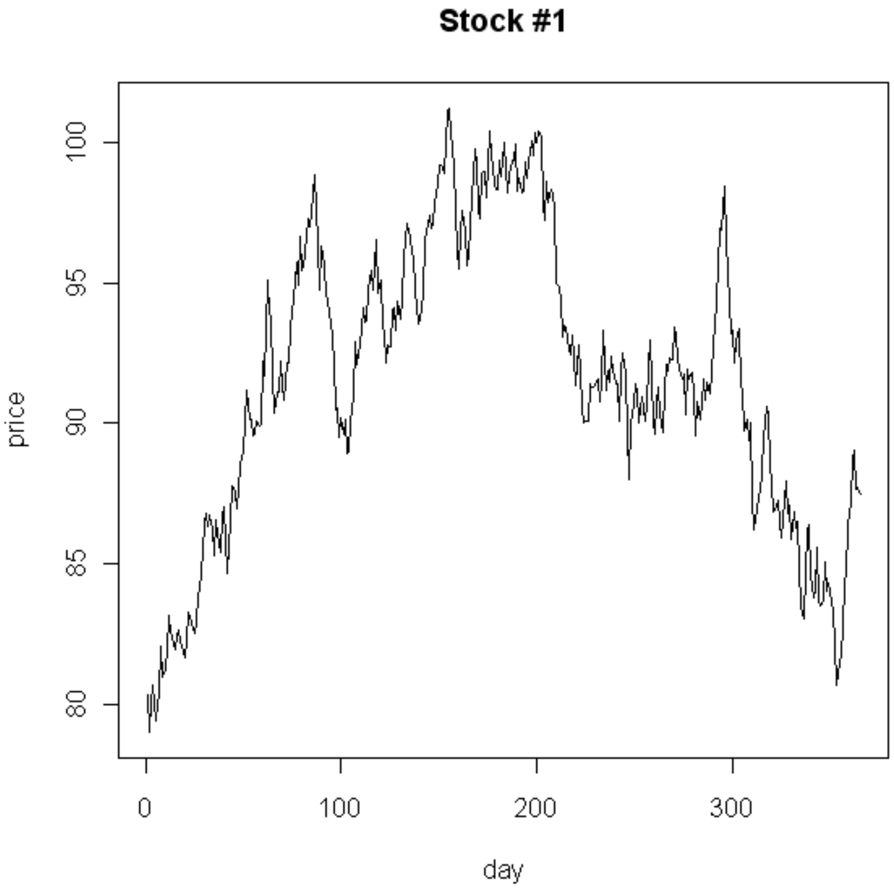
- “When the facts change, I change my mind. What do you do sir?”

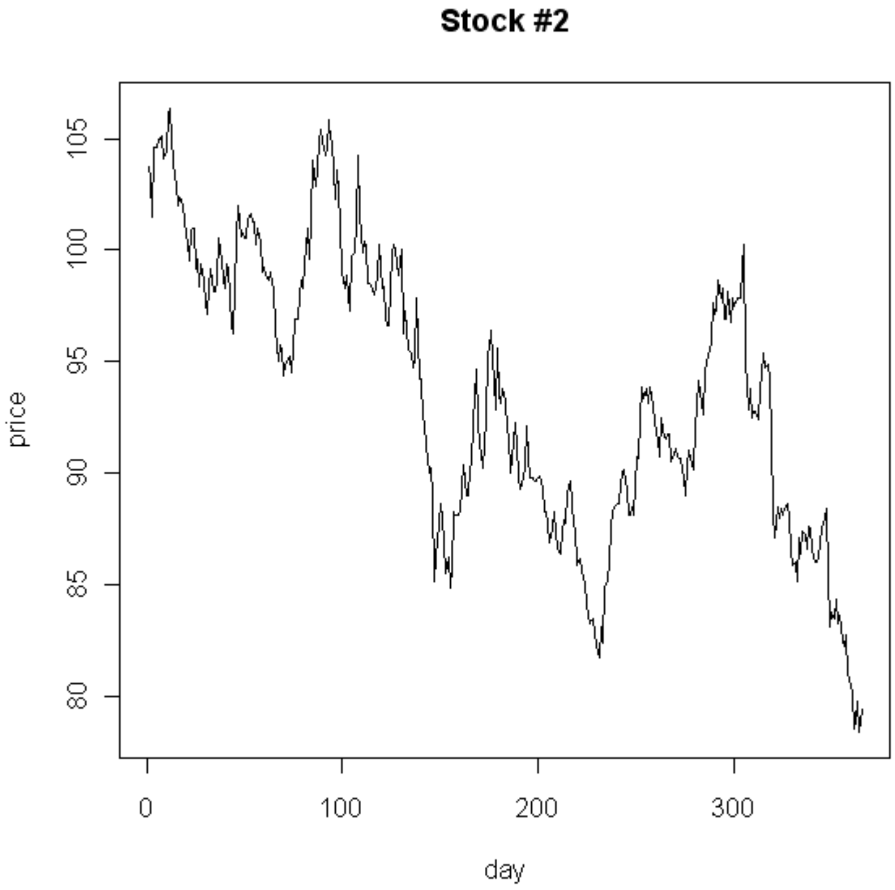
  - John Maynard Keynes

- Variation is everywhere

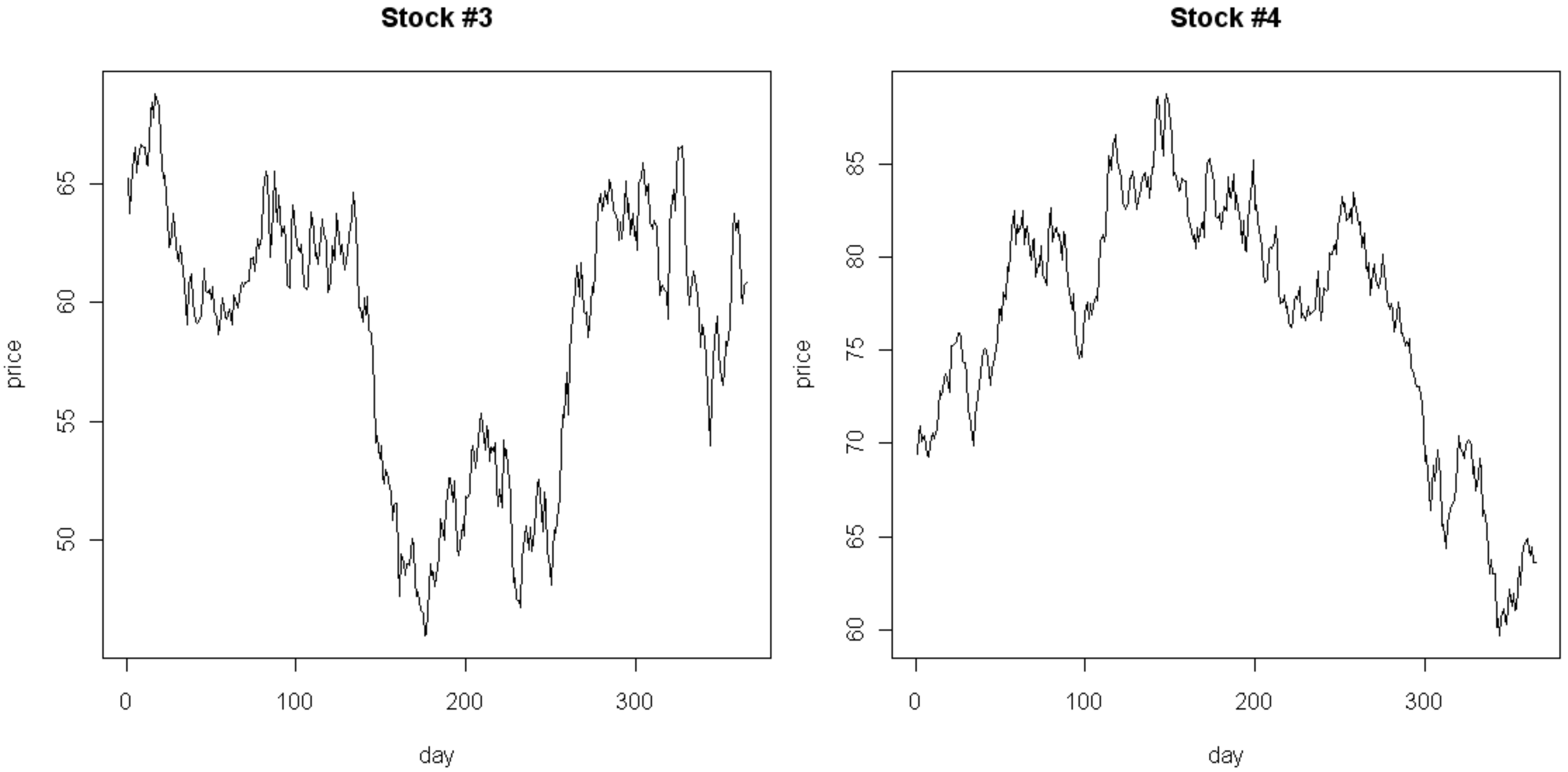
Observed value = Truth + Bias + Random Error

Time Series of four “stock prices”

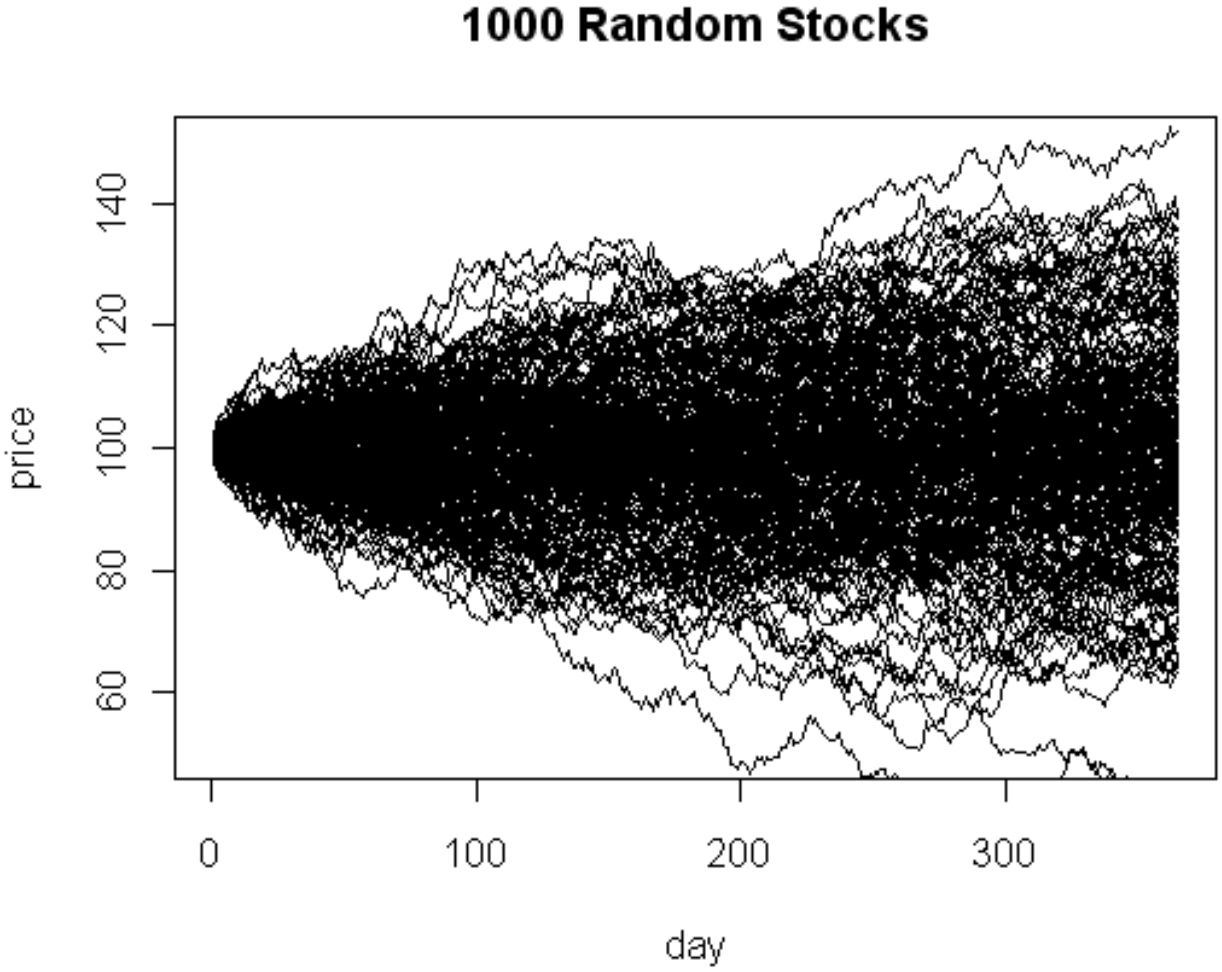




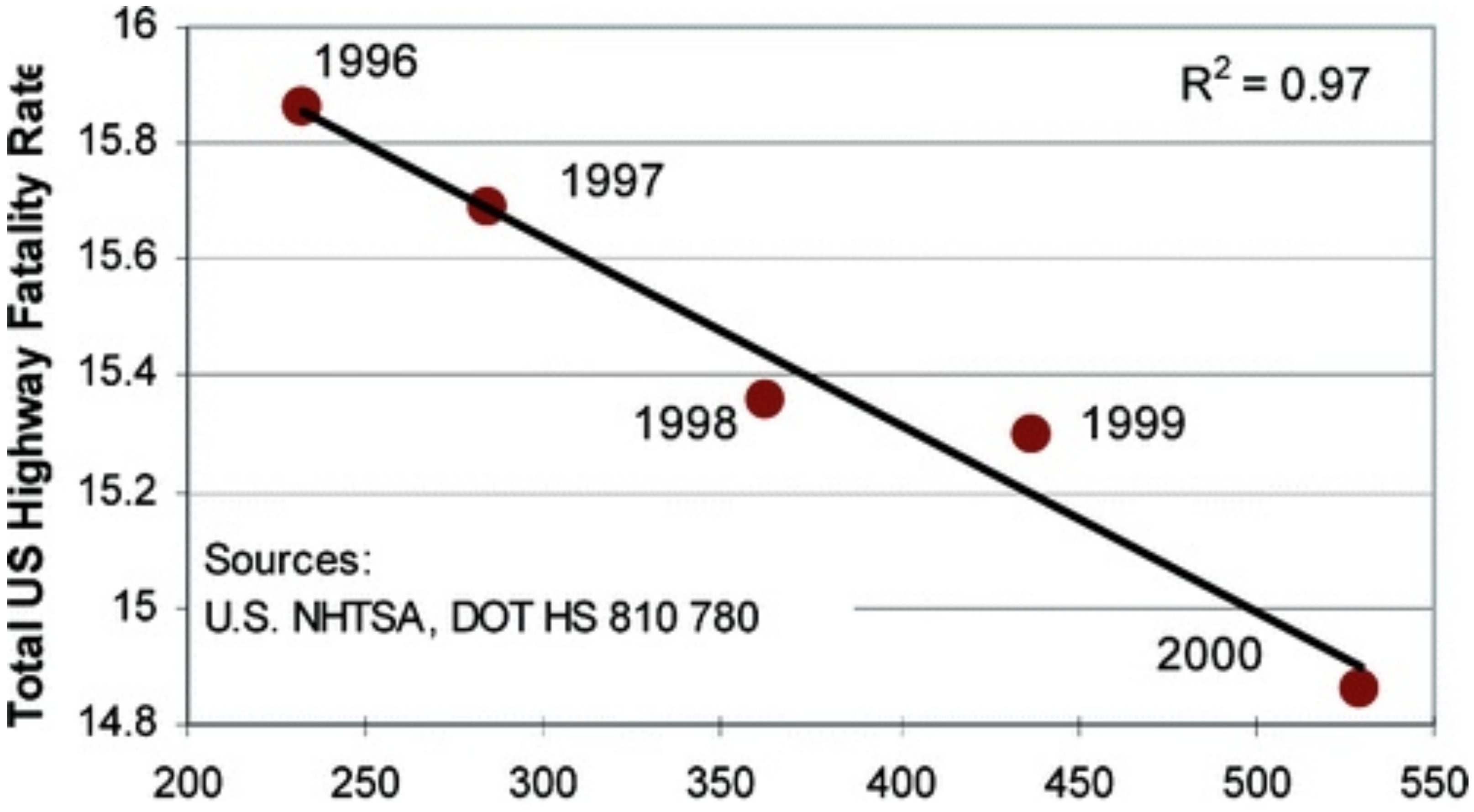
Time series of four “stock prices”



Don’t be fooled by randomness



Confounding



Per Capita Expenditures on Road Maintenance(?)

What is the conclusion?

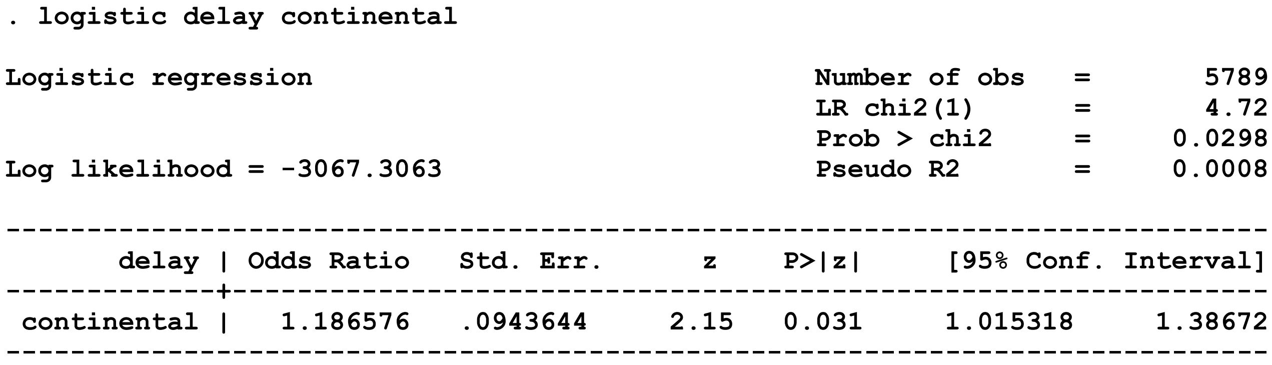
January 2009

Percent of Planes Delayed from City of Origin

	Continental			United		
Airport	Late	Total	%	Late	Total	%
Newark	957	3998	***23.9***	100	399	25.1
LaGuardia	62	356	***17.4***	113	573	19.7
Pittsburg	8	60	***13.3***	17	119	14.3
Detroit	16	145	***11.0***	16	139	11.5
Totals	1043	4559	22.9	246	1230	***20.0***

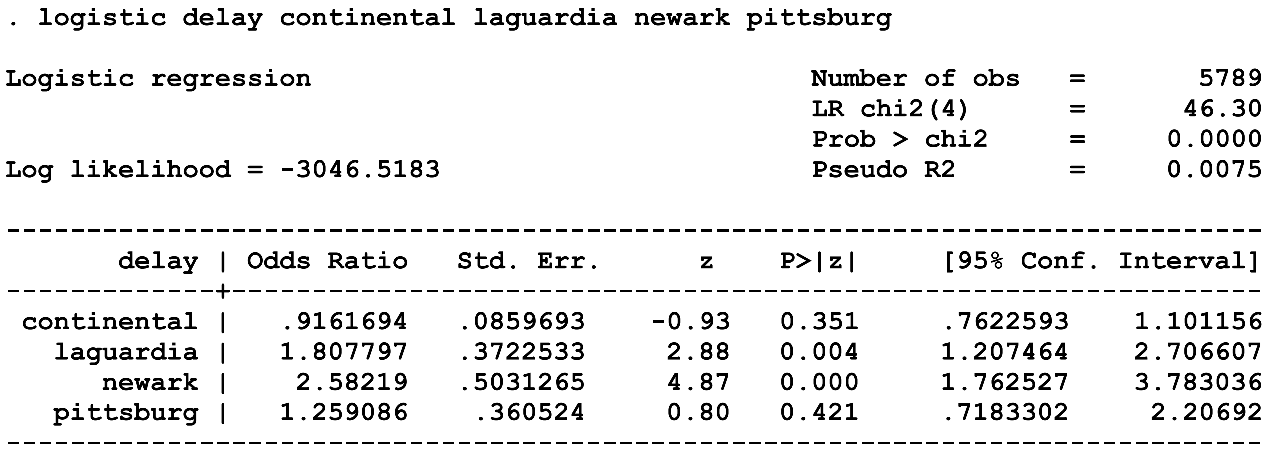
slide credit: Jeff Witmer, data source: www.bts.gov

How about now?



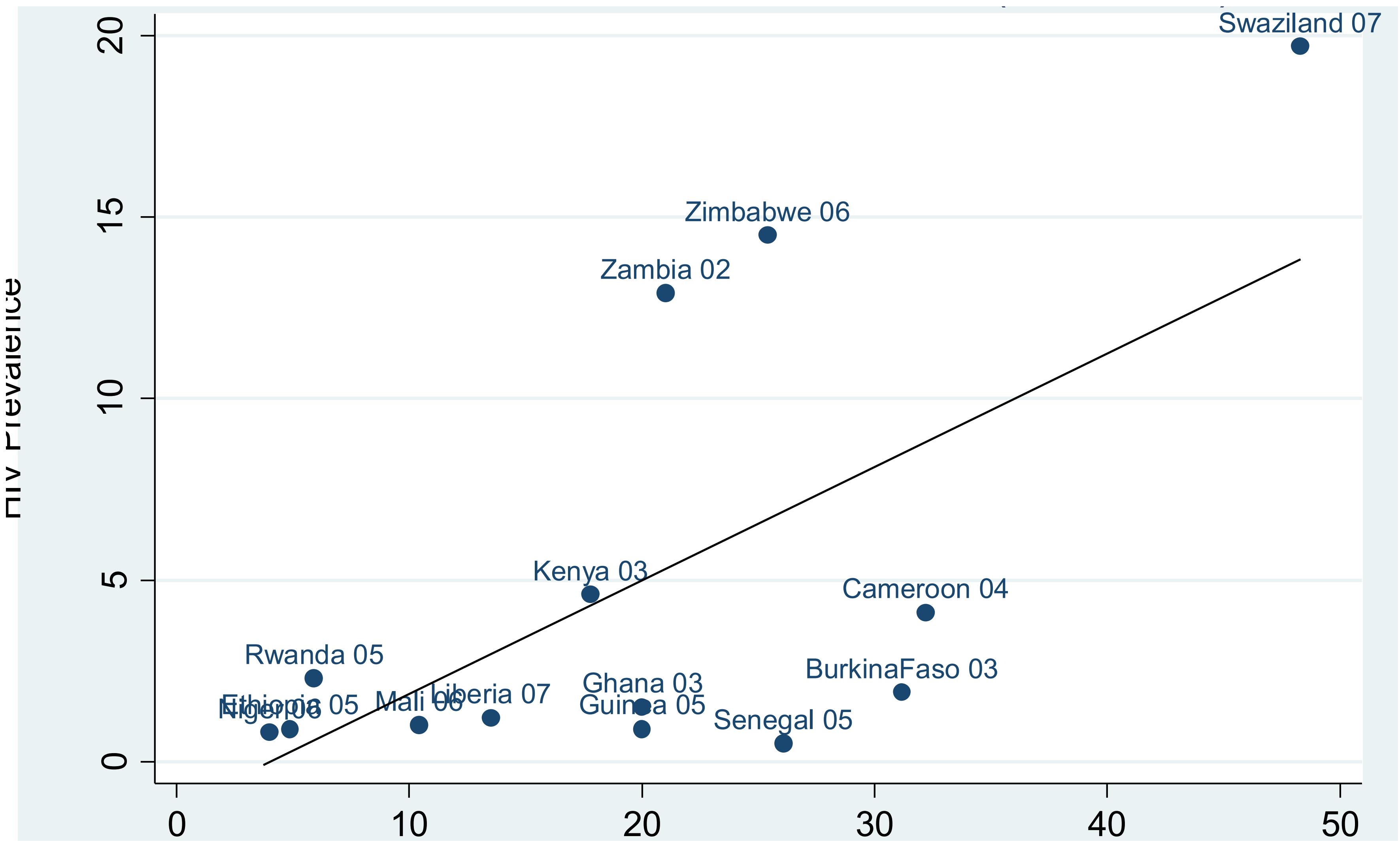
```
. logistic delay continental

Logistic regression                             Number of obs     =       5789
                                                LR chi2(1)        =       4.72
                                                Prob > chi2       =     0.0298
Log likelihood = -3067.3063                     Pseudo R2         =     0.0008

------------------------------------------------------------------------------
       delay | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 continental |   1.186576   .0943644     2.15   0.031     1.015318     1.38672
------------------------------------------------------------------------------
```

Unadjusted
OR = 1.19
95% CI = (1.02, 1.39)

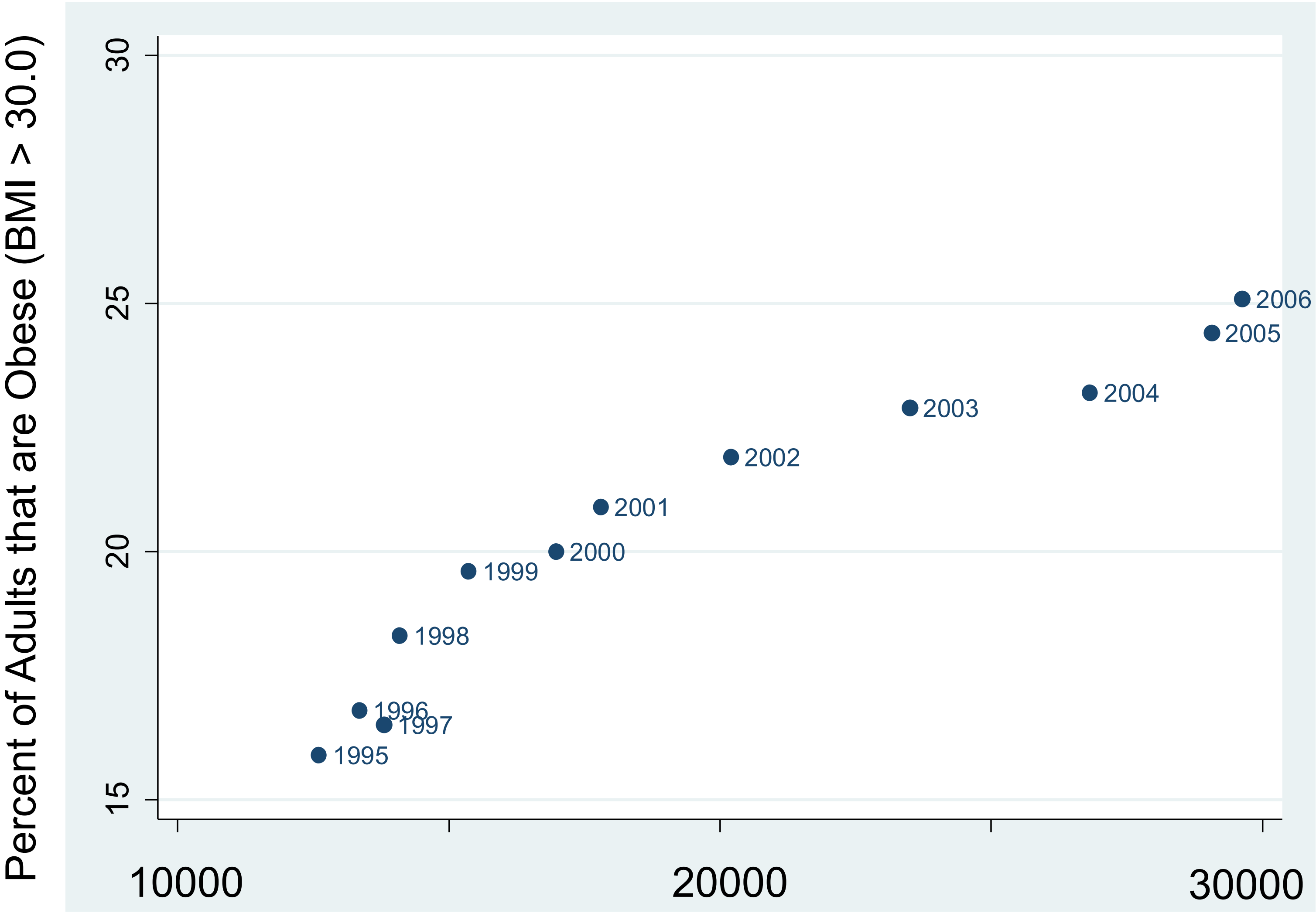


```
. logistic delay continental laguardia newark pittsburg

Logistic regression                             Number of obs     =       5789
                                                LR chi2(4)        =      46.30
                                                Prob > chi2       =     0.0000
Log likelihood = -3046.5183                     Pseudo R2         =     0.0075

------------------------------------------------------------------------------
       delay | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 continental |   .9161694   .0859693    -0.93   0.351     .7622593    1.101156
   laguardia |   1.807797   .3722533     2.88   0.004     1.207464    2.706607
      newark |    2.58219   .5031265     4.87   0.000     1.762527    3.783036
   pittsburg |   1.259086    .360524     0.80   0.421     .7183302     2.20692
------------------------------------------------------------------------------
```

Adjusted
OR = 0.92
95% CI = (0.76, 1.10)


HIV Prevalence
0
5
10
15
20
0
10
20
30
40
50
Swaziland 07
Zimbabwe 06
Zambia 02
Kenya 03
Cameroon 04
Rwanda 05
BurkinaFaso 03
Ghana 03
Guinea 05
Senegal 05
Niger 06
Ethiopia 05
Mali 06
Liberia 07

Obesity in the United States: 1995 - 2006



Number of Gyms in the U.S.

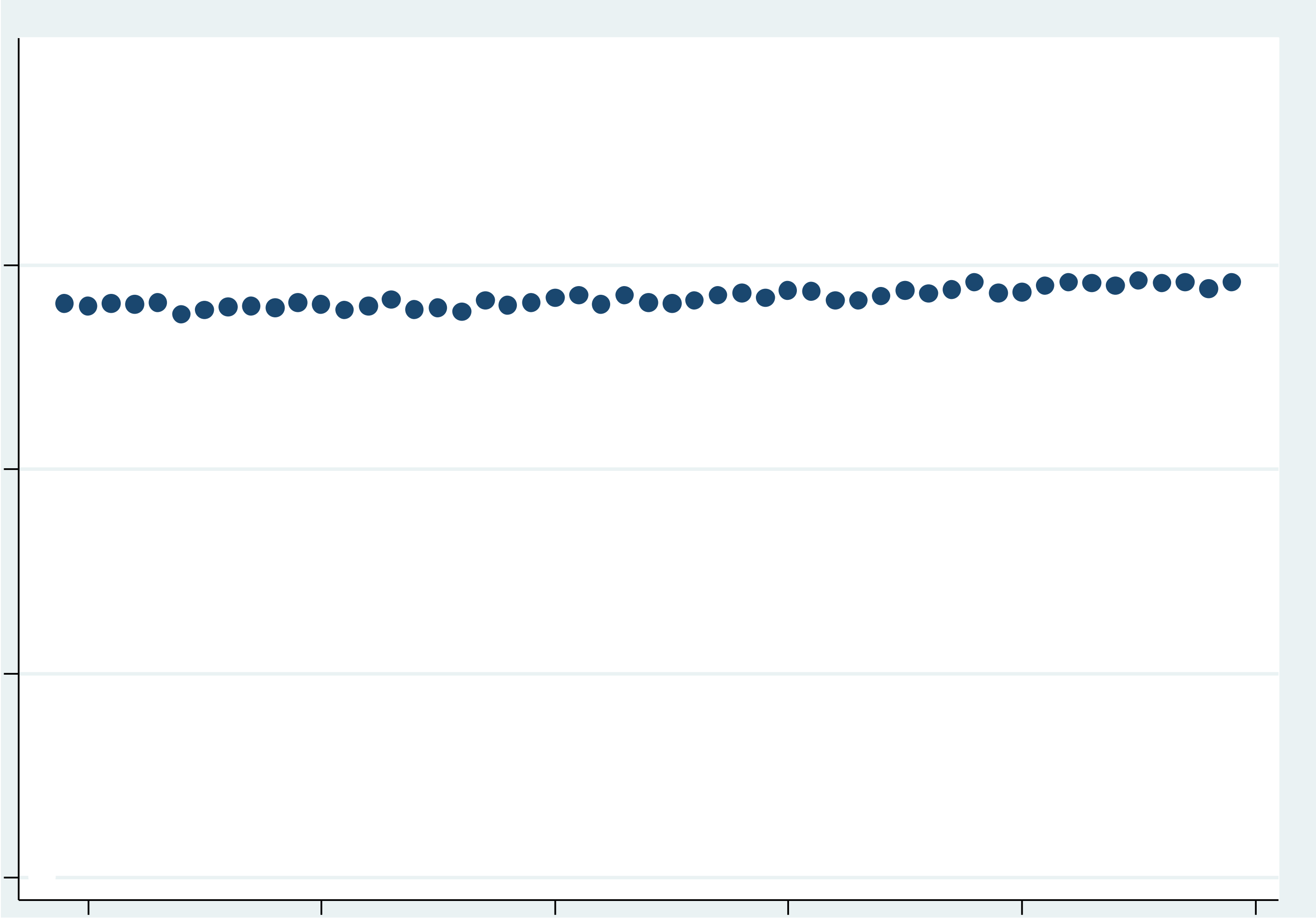
Sources: CDC (BRFSS),
Merriman, Curhan, & Ford Fitness & Wellness report, 2006

What do the data say?

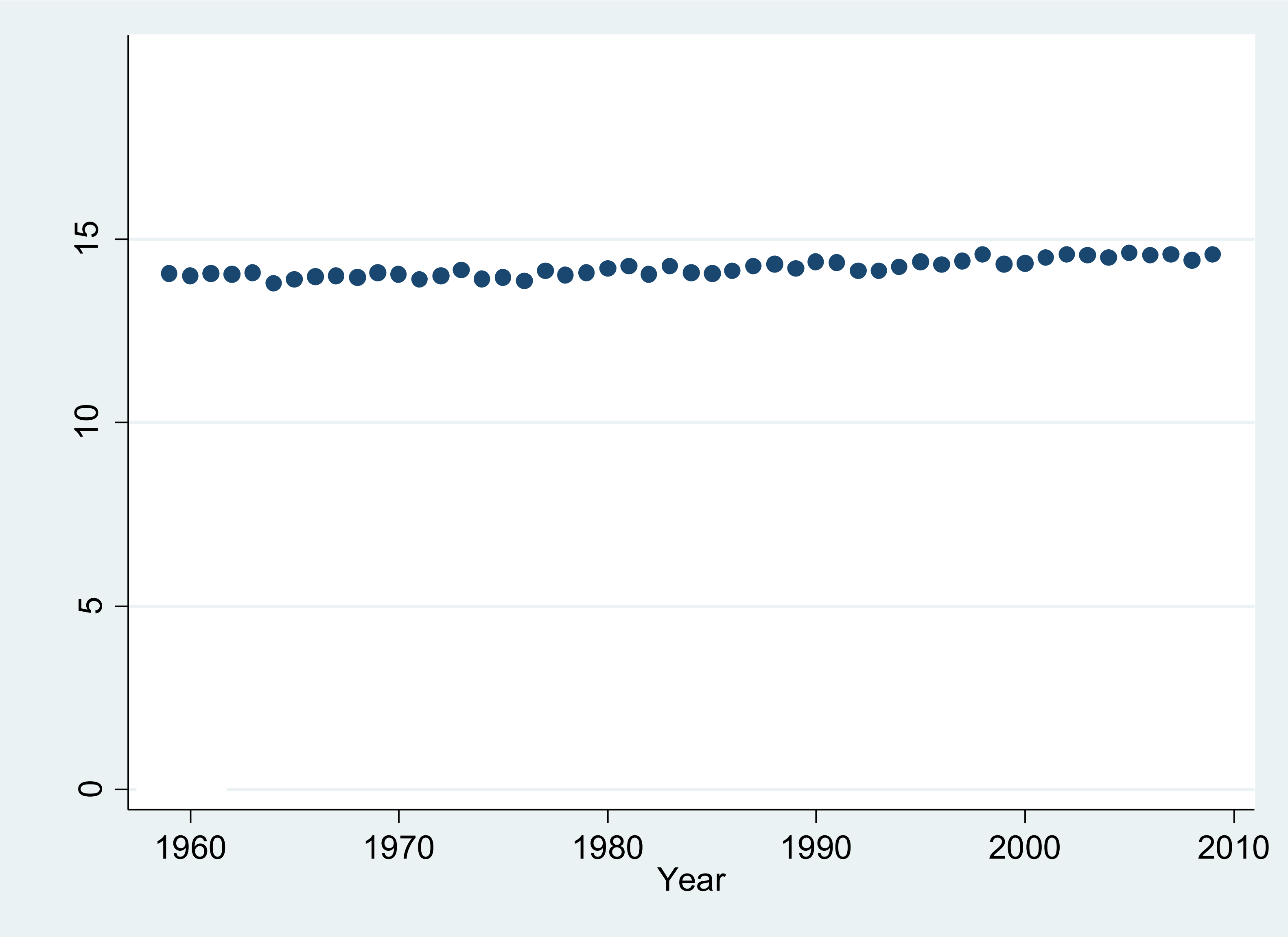
- Caution:

- This may require thoughtful consideration

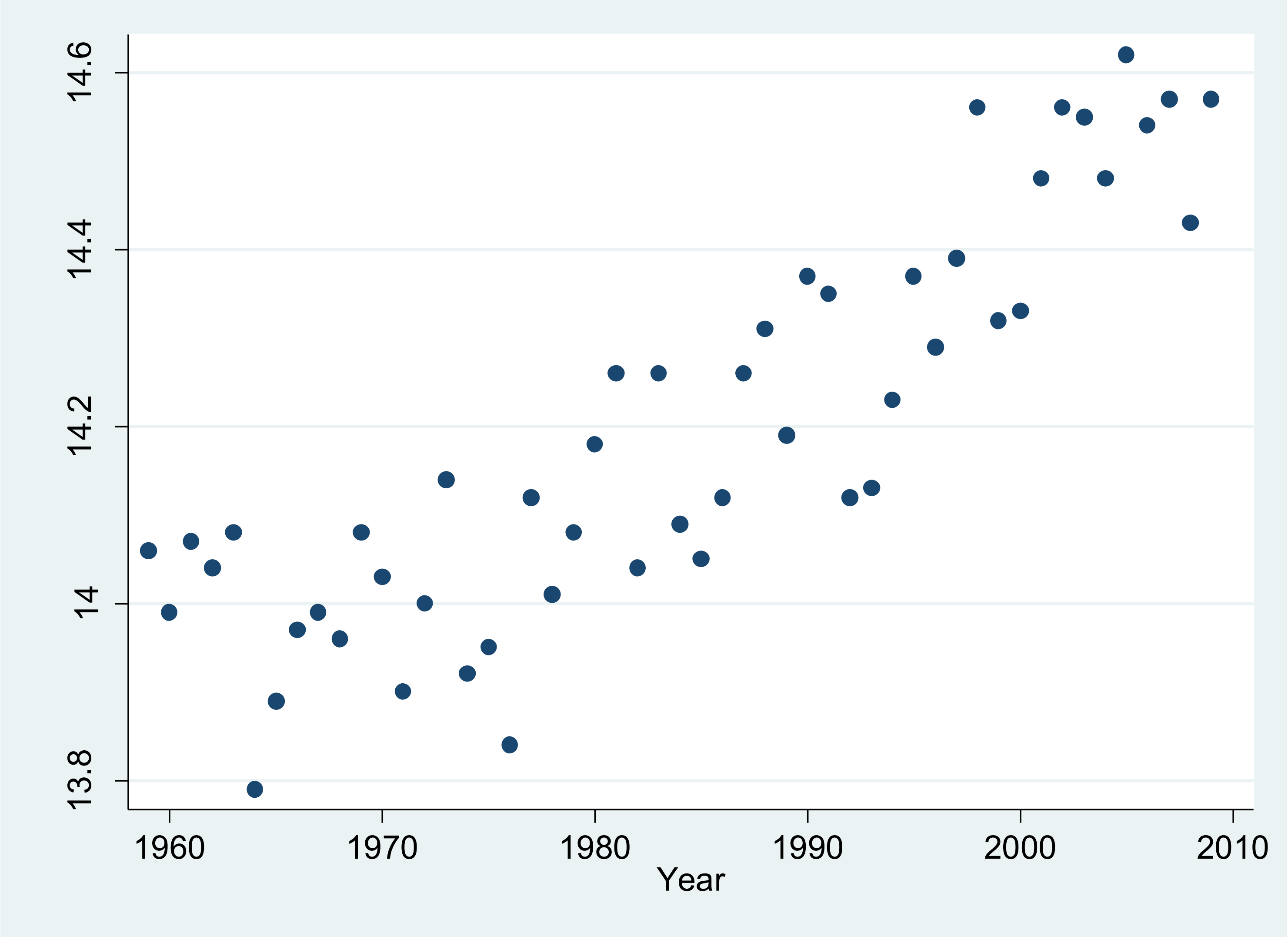
What is the relationship?



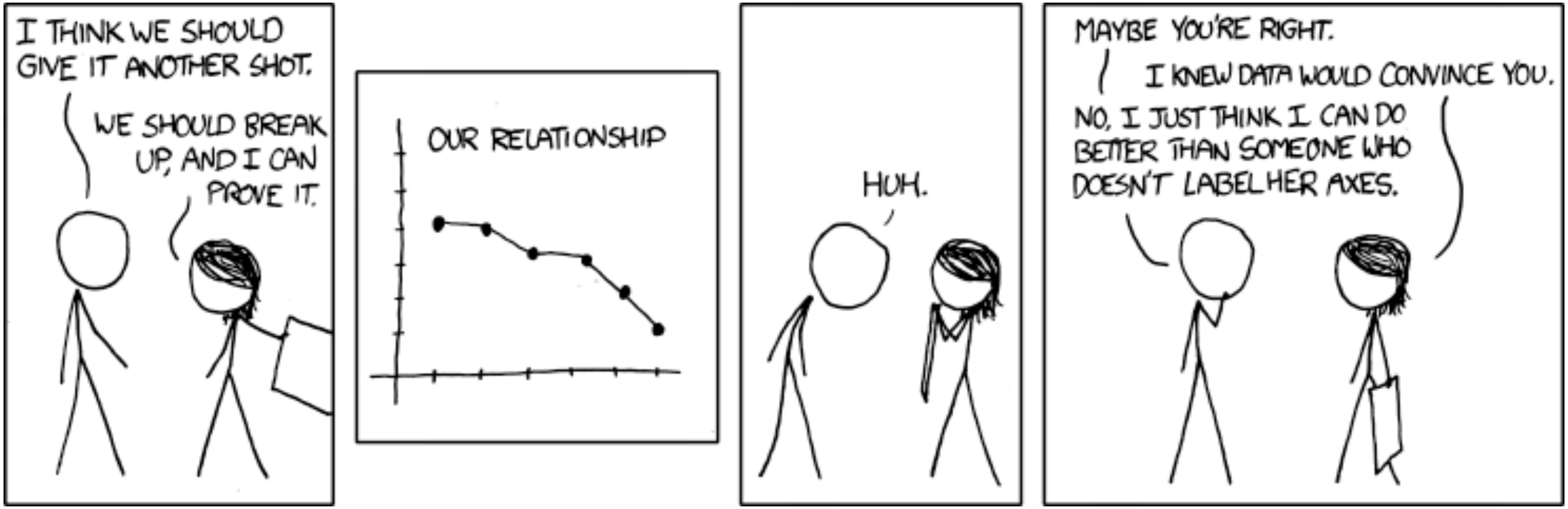
What is the relationship?



Median Global Temperature During the Past 50 Years

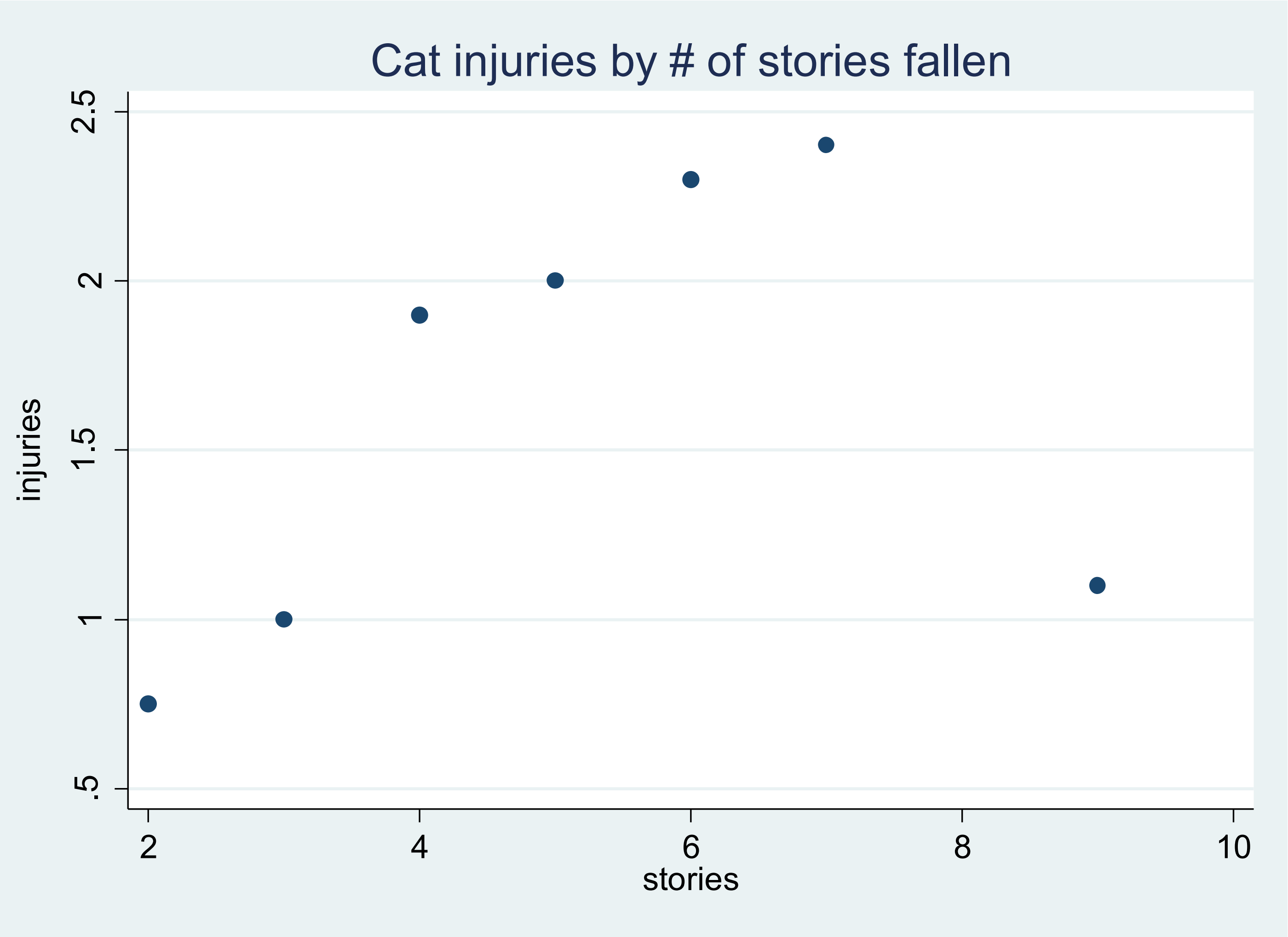


Love statistics(?)


I THINK WE SHOULD GIVE IT ANOTHER SHOT.
WE SHOULD BREAK UP, AND I CAN PROVE IT.
OUR RELATIONSHIP
HUH.
MAYBE YOU'RE RIGHT.
I KNEW DATA WOULD CONVINCE YOU.
NO, I JUST THINK I CAN DO BETTER THAN SOMEONE WHO DOESN'T LABEL HER AXES.

xkcd

Are you a believer?

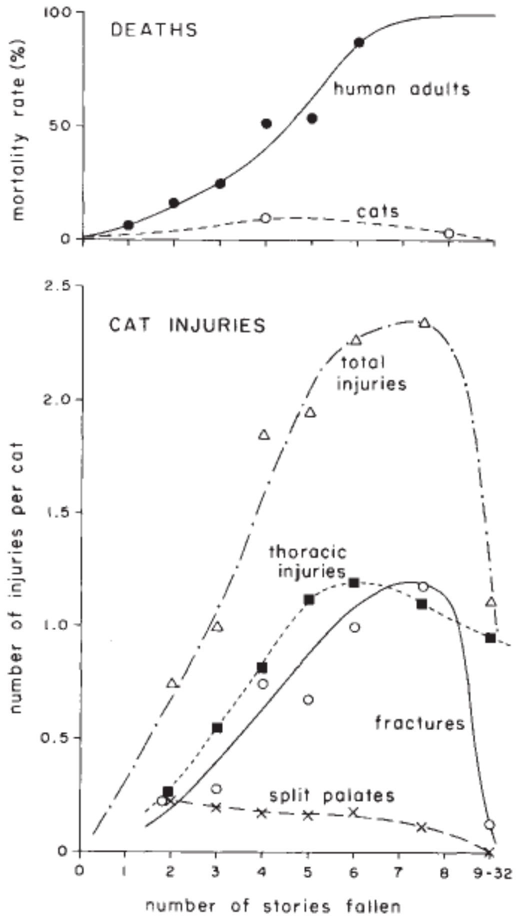


Cat conundrum

As long as it experiences acceleration, the cat probably extends its limbs reflexly, but on reaching terminal velocity it may relax and extend the limbs more horizontally in flying-squirrel fashion, thus not only reducing the velocity of fall but also absorbing the impact over a greater area of its body. This may explain the paradoxical decrease of mortality and injury in cats that fall more than 100 feet.

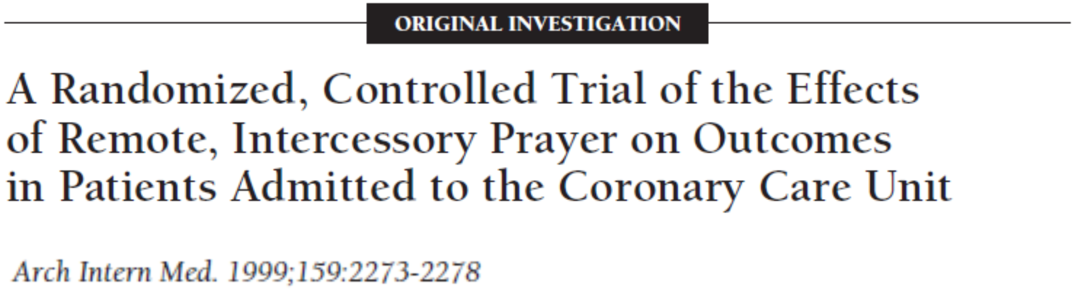
Stories Fallen	# of cats
1	0
2	8
3	14
4	27
5	34
6	21
7-8	9
9-32	13

NATURE VOL. 332 14 APRIL 1988



Mortality rates for falling adult humans and cats (above), and number of total injuries and various types of injury per falling cat (below), as a function of number of stories fallen. (Based on the work by Waring and Demling and by Whitney and Mehlhoff.)

What do you think?



ORIGINAL INVESTIGATION

A Randomized, Controlled Trial of the Effects of Remote, Intercessory Prayer on Outcomes in Patients Admitted to the Coronary Care Unit

Arch Intern Med. 1999;159:2273-2278

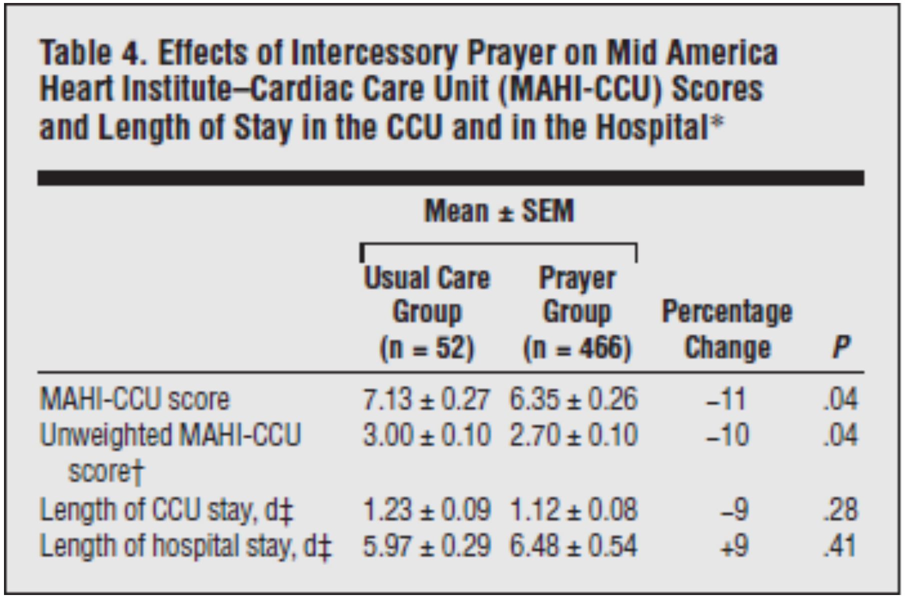


Table 4. Effects of Intercessory Prayer on Mid America Heart Institute–Cardiac Care Unit (MAHI-CCU) Scores and Length of Stay in the CCU and in the Hospital*

	Mean ± SEM			
	Usual Care Group (n = 52)	Prayer Group (n = 466)	Percentage Change	*P*
MAHI-CCU score	7.13 ± 0.27	6.35 ± 0.26	−11	.04
Unweighted MAHI-CCU score†	3.00 ± 0.10	2.70 ± 0.10	−10	.04
Length of CCU stay, d‡	1.23 ± 0.09	1.12 ± 0.08	−9	.28
Length of hospital stay, d‡	5.97 ± 0.29	6.48 ± 0.54	+9	.41

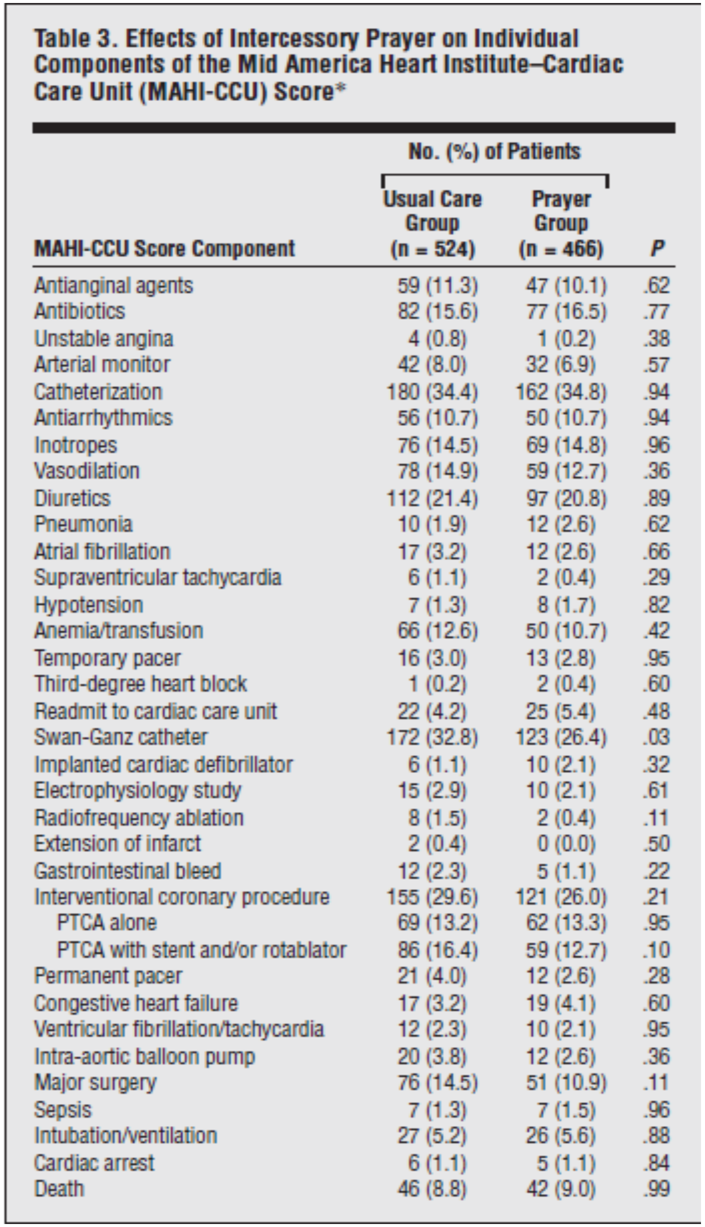
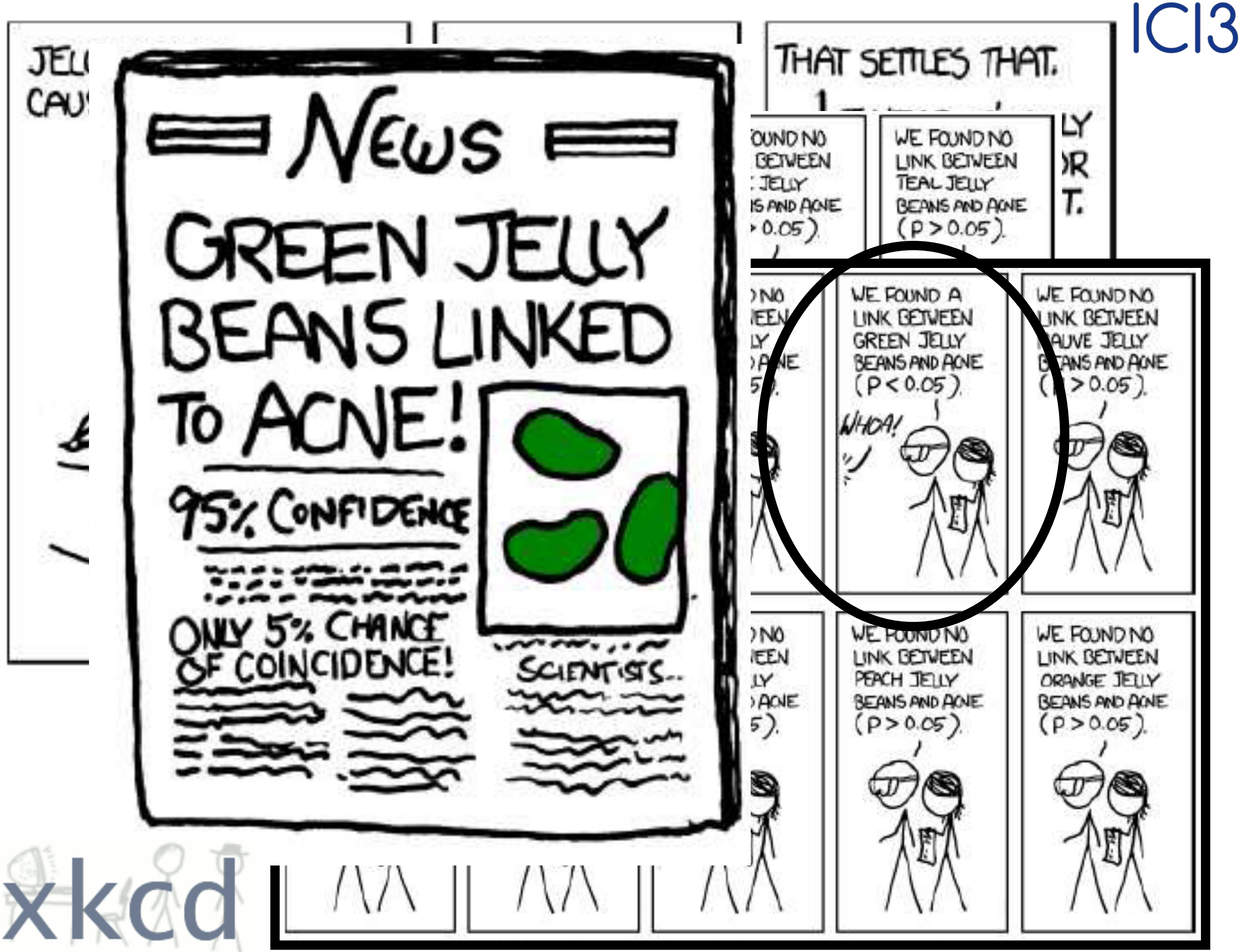


Table 3. Effects of Intercessory Prayer on Individual Components of the Mid America Heart Institute–Cardiac Care Unit (MAHI-CCU) Score*

	No. (%) of Patients		
MAHI-CCU Score Component	Usual Care Group (n = 524)	Prayer Group (n = 466)	*P*
Antianginal agents	59 (11.3)	47 (10.1)	.62
Antibiotics	82 (15.6)	77 (16.5)	.77
Unstable angina	4 (0.8)	1 (0.2)	.38
Arterial monitor	42 (8.0)	32 (6.9)	.57
Catheterization	180 (34.4)	162 (34.8)	.94
Antiarrhythmics	56 (10.7)	50 (10.7)	.94
Inotropes	76 (14.5)	69 (14.8)	.96
Vasodilation	78 (14.9)	59 (12.7)	.36
Diuretics	112 (21.4)	97 (20.8)	.89
Pneumonia	10 (1.9)	12 (2.6)	.62
Atrial fibrillation	17 (3.2)	12 (2.6)	.66
Supraventricular tachycardia	6 (1.1)	2 (0.4)	.29
Hypotension	7 (1.3)	8 (1.7)	.82
Anemia/transfusion	66 (12.6)	50 (10.7)	.42
Temporary pacer	16 (3.0)	13 (2.8)	.95
Third-degree heart block	1 (0.2)	2 (0.4)	.60
Readmit to cardiac care unit	22 (4.2)	25 (5.4)	.48
Swan-Ganz catheter	172 (32.8)	123 (26.4)	.03
Implanted cardiac defibrillator	6 (1.1)	10 (2.1)	.32
Electrophysiology study	15 (2.9)	10 (2.1)	.61
Radiofrequency ablation	8 (1.5)	2 (0.4)	.11
Extension of infarct	2 (0.4)	0 (0.0)	.50
Gastrointestinal bleed	12 (2.3)	5 (1.1)	.22
Interventional coronary procedure	155 (29.6)	121 (26.0)	.21
PTCA alone	69 (13.2)	62 (13.3)	.95
PTCA with stent and/or rotablator	86 (16.4)	59 (12.7)	.10
Permanent pacer	21 (4.0)	12 (2.6)	.28
Congestive heart failure	17 (3.2)	19 (4.1)	.60
Ventricular fibrillation/tachycardia	12 (2.3)	10 (2.1)	.95
Intra-aortic balloon pump	20 (3.8)	12 (2.6)	.36
Major surgery	76 (14.5)	51 (10.9)	.11
Sepsis	7 (1.3)	7 (1.5)	.96
Intubation/ventilation	27 (5.2)	26 (5.6)	.88
Cardiac arrest	6 (1.1)	5 (1.1)	.84
Death	46 (8.8)	42 (9.0)	.99


JELL
CAU
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...
THAT SETTLES THAT.
LY
OR
T.
OUND NO
BETWEEN
JELLY
S AND ACNE
0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
NO
EEN
LY
ACNE
5).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
NO
EEN
LY
ACNE
5).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
xkcd

Skepticism vs Openness

- "It seems to me what is called for is an exquisite balance between two conflicting needs: the most skeptical scrutiny of all hypotheses that are served up to us and at the same time a great openness to new ideas …
- If you are only skeptical, then no new ideas make it through to you …
- On the other hand, if you are open to the point of gullibility and have not an ounce of skeptical sense in you, then you cannot distinguish the useful ideas from the worthless ones."

- Carl Sagan

Summary

- We use data to help form and revise models
- Data collection should not be haphazard
- Data are not easily obtained – especially good data
- Think about how the data were collected and be sure to consider:
    - Bias and confounding
    - Variability
    - How the data are presented
- Be skeptical but open

Title: Thinking About Data. DOI: 10.6084/m9.figshare.5044615.v4

Attribution:
Jim Scott

Clinic on the Meaningful Modeling of Epidemiological Data

Source URL:
https://doi.org/10.6084/m9.figshare.5044615.v4

For further information please contact figshare@ici3d.org.

AIMS African Institute for Mathematical Sciences SOUTH AFRICA
SACEMA DST-NRF Centre of Excellence in Epidemiological Modelling and Analysis
UNIVERSITY OF GEORGIA College of Public Health