Health Informatics

Lecture 7

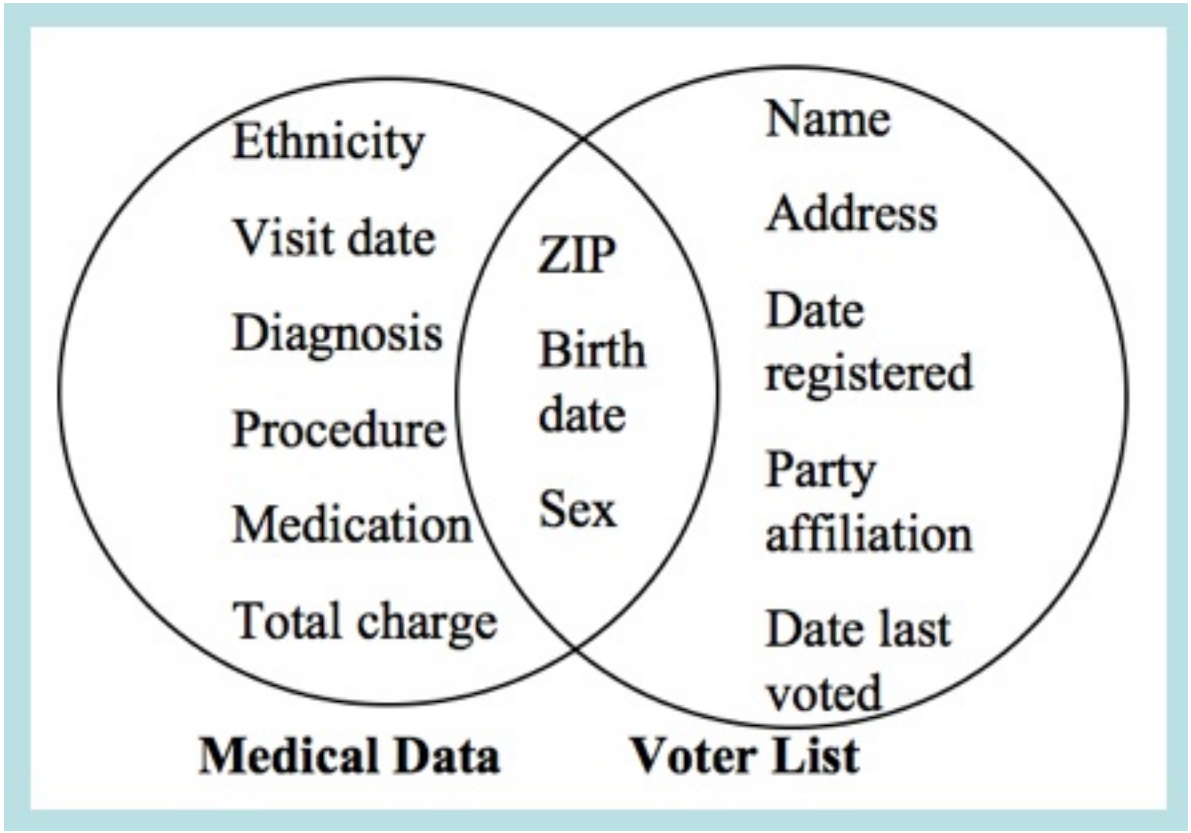
Samantha Kleinberg
samantha.kleinberg@stevens.edu

Practicalities: preserving privacy

Good practices for preserving privacy

- Physical security
  - Locked cabinet/office, don't print records and leave them in common area/on desk
- Encryption
  - Losing laptop/thumbdrive, sending harddrive for repair can be disastrous
- Restricted access to servers with PHI
- PHI in segregated, labeled, directories

Limits of deidentification



"87% of the U.S. Population are uniquely identified by {date of birth, gender, ZIP}."

http://latanyasweeney.org/work/identifiability.html

Matching Known Patients to Health Records in Washington State Data

Latanya Sweeney

(Submitted on 4 Jul 2013 (v1), last revised 5 Jul 2013 (this version, v2))

The State of Washington sells patient-level health data for $50. This publicly available dataset has virtually all hospitalizations occurring in the State in a given year, including patient demographics, diagnoses, procedures, attending physician, hospital, a summary of charges, and how the bill was paid. It does not contain patient names or addresses (only ZIPs). Newspaper stories printed in the State for the same year that contain the word "hospitalized" often include a patient's name and residential information and explain why the person was hospitalized, such as vehicle accident or assault. News information uniquely and exactly matched medical records in the State database for 35 of the 81 cases (or 43 percent) found in 2011, thereby putting names to patient records. A news reporter verified matches by contacting patients. Employers, financial organizations and others know the same kind of information as reported in news stories making it just as easy for them to identify the medical records of employees, debtors, and others.

Comments: 12 pages

Name of Covered Entity	State	Individuals Affected	Breach Date	Type of Breach	Location of Breached Information
Supportive Concepts For Families, Inc.	PA	593	02/06/2013 - 12/16/2013	Unauthorized Access/Disclosure	Network Server
Health Care Solutions At Home Inc.	OH	1139	12/17/2013	Other	Other
St. Vincent Hospital And Healthcare, Inc	IN	1142	12/23/2013	Theft	Laptop
Missouri Consolidated Health Care Plan	MO	10024	03/23/2012 - 01/22/2014	Unauthorized Access/Disclosure	Network Server
The Clorox Company Group Insurance Plan	CA	520	04/16/2012 - 01/22/2014	Unauthorized Access/Disclosure	Network Server
Inspira Health Network Inc.	NJ	1411	12/23/2013 - 12/23/2013	Theft	Desktop Computer
Nissan North America, Inc.	TN	1511	05/08/2012 - 01/22/2014	Unauthorized Access/Disclosure	Network Server
Health Source Of Ohio Inc.	OH	8845	11/18/2013 - 12/24/2013	Unauthorized Access/Disclosure, Other	Other
Cornerstone Health Care, Pa	NC	548	12/31/2013	Theft, Loss	Laptop

The False Security of Blind Dates

Chrononymization's Lack of Impact on Data Privacy of Laboratory Data

J.J. Cimino

NIH Clinical Center, US National Institutes of Health, Bethesda, Maryland, USA

Keywords

Patient data privacy, data adjustments, clinical research, clinical informatics, health policy, anonymizatoin, de-identification, dates

Summary

Background: The reuse of clinical data for research purposes requires methods for the protection of personal privacy. One general approach is the removal of personal identifiers from the data. A frequent part of this anonymization process is the removal of times and dates, which we refer to as "chrononymization." While this step can make the association with identified data (such as public information or a small sample of patient information) more difficult, it comes at a cost to the usefulness of the data for research.

Objectives: We sought to determine whether removal of dates from common laboratory test panels offers any advantage in protecting such data from re-identification.

Methods: We obtained a set of results for 5.9 million laboratory panels from the National Institutes of Health's (NIH) Biomedical Translational Research Information System (BTRIS), selected a random set of 20,000 panels from the larger source sets, and then identified all matches between the sets.

Results: We found that while removal of dates could hinder the re-identification of a single test result, such removal had almost no effect when entire panels were used.

Conclusions: Our results suggest that reliance on chrononymization provides a false sense of security for the protection of laboratory test results. As a result of this study, the NIH has chosen to rely on policy solutions, such as strong data use agreements, rather than removal of dates when reusing clinical data for research purposes.

Back to data challenges

- Observational Data in general
  - Nonstationarity
  - Selection bias
  - Choosing variables
  - Seeing vs doing
- Biomedical Data
  - Biased approximation of truth
  - Sample size
  - Censoring (left, right)
  - Institutional differences
  - Fragmentation of data

Treatment		
	Dead	Alive
A	85	215 (72%)
B	59	241 (80%)
Total	144	456

Simpson's paradox

Treatment	Men		Women		Combined	
	Dead	Alive	Dead	Alive	Dead	Alive
A	80	120 (60%)	5	95 (95%)	85	215 (72%)
B	20	20 (50%)	39	221 (85%)	59	241 (80%)
Total	100	140	44	316	144	456

Baker SG, Kramer BS (2001) Good for women, good for men, bad for people: Simpson's paradox and the importance of sex-specific analysis in observational studies. Journal of women's health & gender-based medicine 10: 867-872

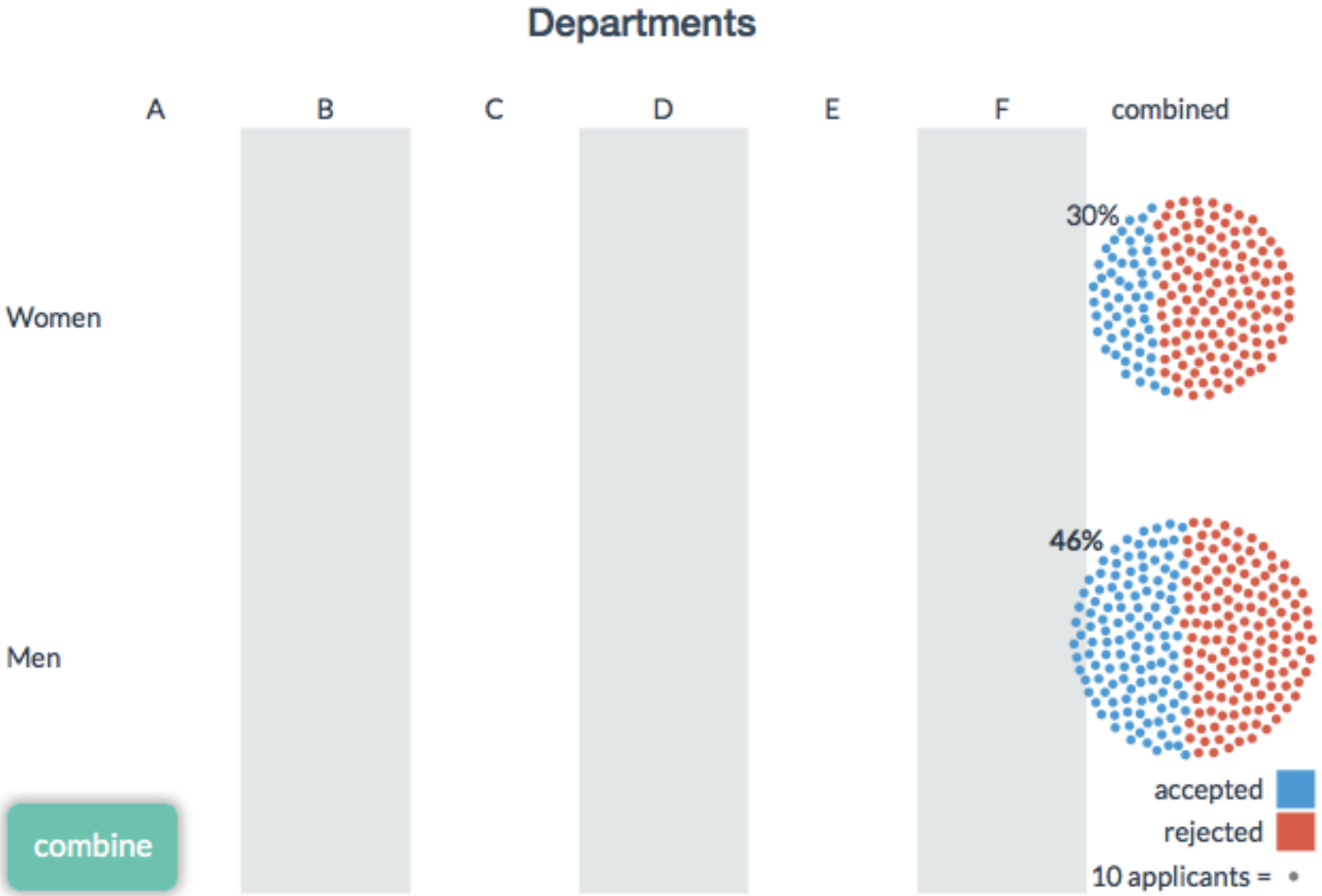
Bias in graduate admissions?

	Admit	Deny
Male	3738	4704
Female	1494	2827

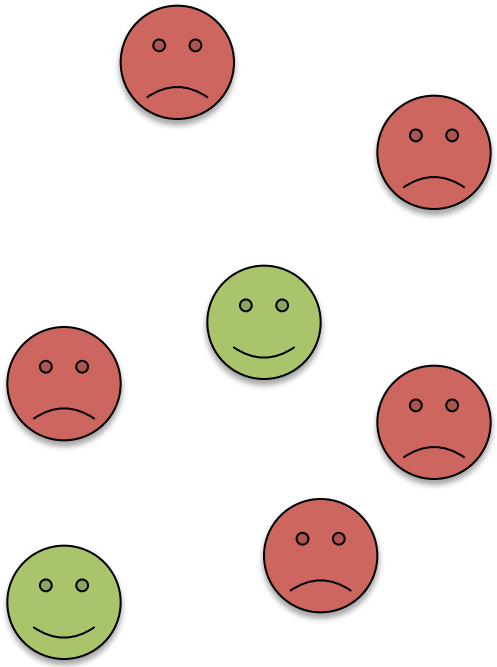
P(admitted)= 5232/12763≈0.41

P(admitted|female) =1494/4321≈0.35
P(admitted|male)=3738/8442≈0.44

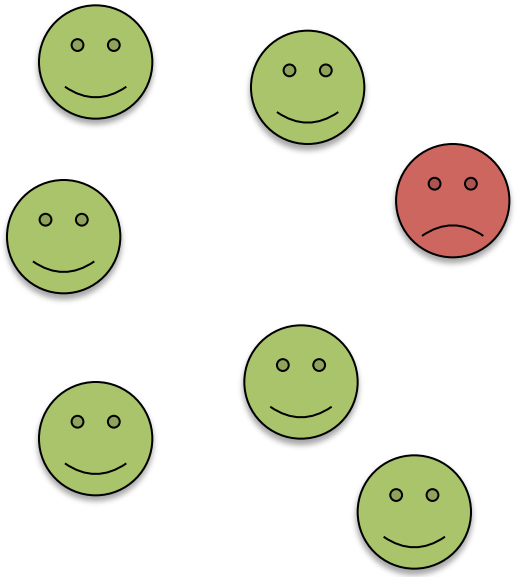
Bickel, P. J., Hammel, E. A., & O'connell, J. W. (1975). Sex bias in graduate admissions: Data from berkeley. *Science*, *187*(4175), 398



http://vudlab.com/simpsons/

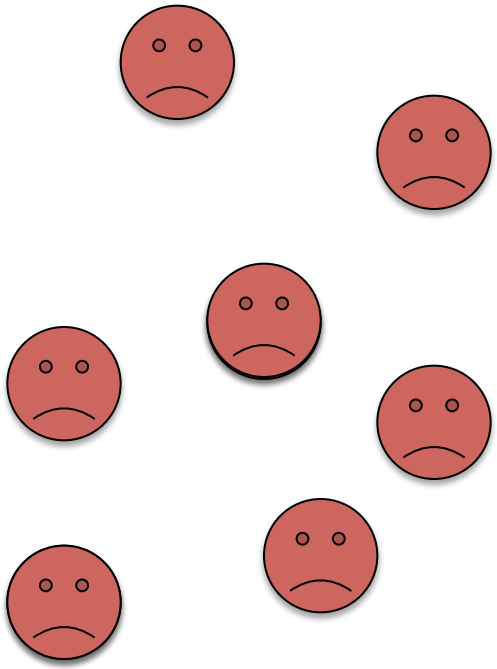


Alice's patient outcomes

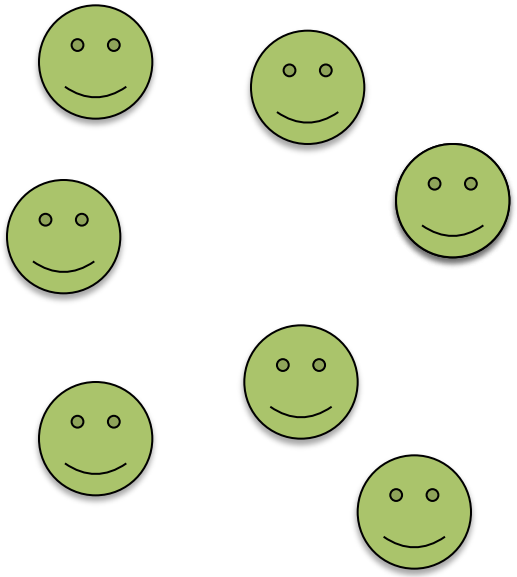


Bob's patient outcomes

Selection bias



Alice's patient population



Bob's patient population

A+H → Hospital

Find association between A, H because not seeing A and H separately in hospital population

Selection bias

- Drug given to sicker patients -> worse outcomes?
- Running/mortality
- Hospital outcomes depend on population: socioeconomic factors, medical insurance, etc.

Variable selection/granularity

- Problem specific
  - What level in ICD9 hierarchy to use?
- Redundancy?
- Hidden common causes

- Nurses' Health study followed ~122,000 nurses every 2 years since the 1970's
- Analysis in the 90's showed women taking HRT after menopause have decreased risk of heart attacks (37% lower death rate, 53% lower risk of CV death)
- HERS trial: RCT showing no effect
- WHI: RCT where heart attacks increase 29% (from 30 to 37 per 10,000 person-years)
- Latest: HRT may be beneficial if it's started early

Doing vs seeing: randomization (more next week)

- Sever link between causes of intervention and effects (selection bias)
  - E.g. birth control pills and pregnancy
- Isolate cause
  - Single difference between groups, removes confounding
- Blinding
  - Confirmation bias

Missing data/error

- Device malfunction
  - E.g. thermometer moves and now measure room temp
- Network problem
  - Data recorded but not stored
- Data exists… somewhere
- Human error – misrecorded data/omitted data
- Patient factors – omitting data/giving false info

Timing variability

- Note describes chronology after the fact
- Labs not exact
- When do people go to dr? (hint: after symptoms start)
- Interventions recorded after they're done

Sample size

- Big data doesn't guarantee sufficient power!

- Would you rather have 10 patients monitored in great detail or a few datapoints on 10K patients?

Censoring

- Left: what happened before admission to hospital?
- Right: what's outcome after leaving hospital?

Fragmentation

- Do you go to the same doctors at home and when classes are in session?
- How are records shared?

Controls

- Usually matched to cases, but technical difficulties
  - Is selection criteria structured or unstructured?
  - Is same data available on both?
  - If comparing against “healthy” people, will there be enough data?
- What happens if some cases are actually controls?

Using Empiric Semantic Correlation to Interpret Temporal Assertions in Clinical Texts

George Hripcsak, MD, MS, Noémie Elhadad, PhD, Yueh-Hsia Chen, MS, Li Zhou, BMed, PhD, Frances P. Morrison, MD, MPH

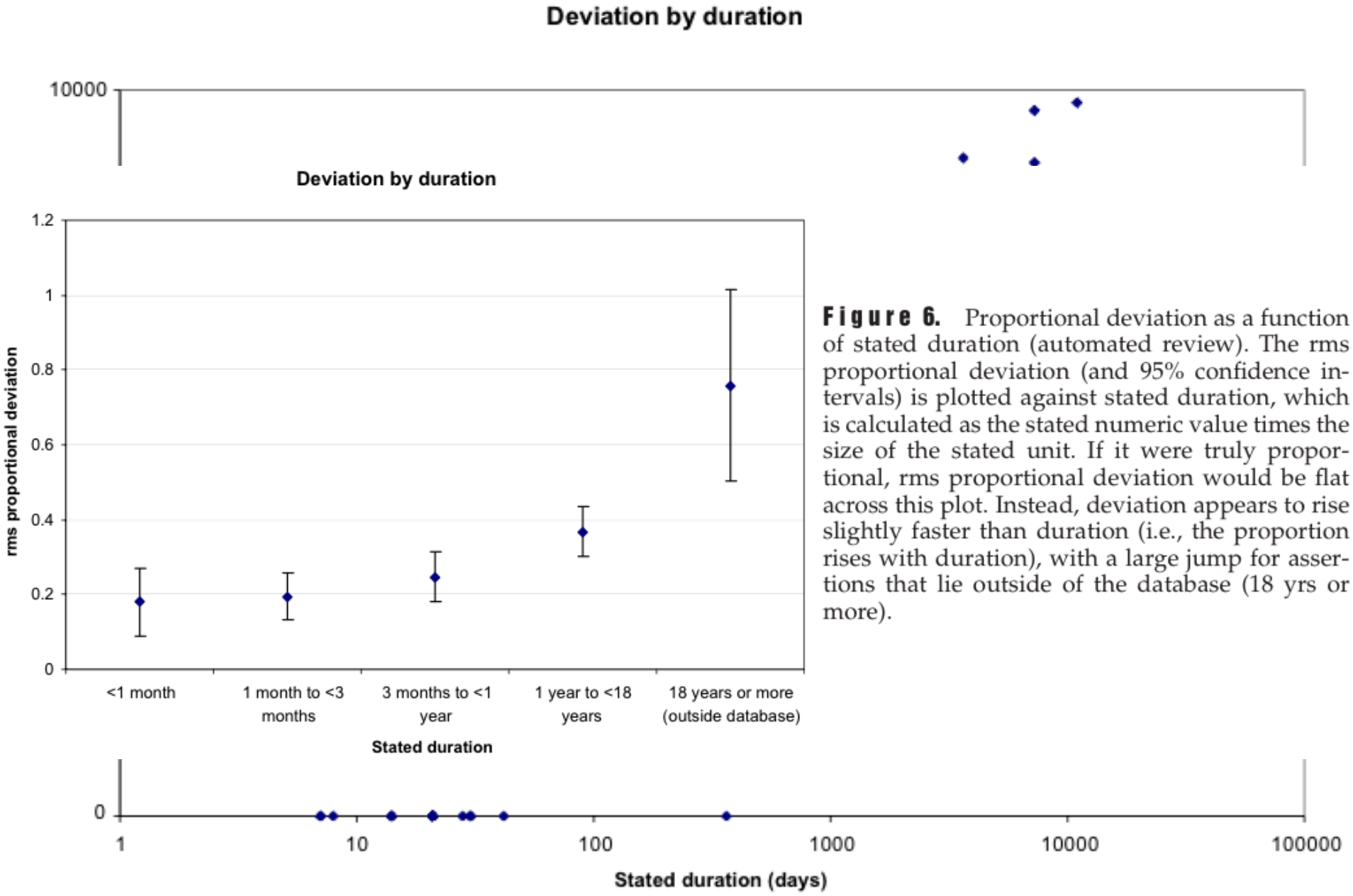


Figure 6. Proportional deviation as a function of stated duration (automated review). The rms proportional deviation (and 95% confidence intervals) is plotted against stated duration, which is calculated as the stated numeric value times the size of the stated unit. If it were truly proportional, rms proportional deviation would be flat across this plot. Instead, deviation appears to rise slightly faster than duration (i.e., the proportion rises with duration), with a large jump for assertions that lie outside of the database (18 yrs or more).

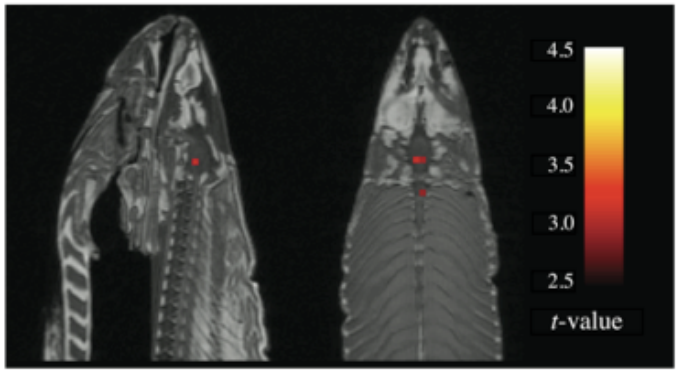
Multiple Testing

METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.



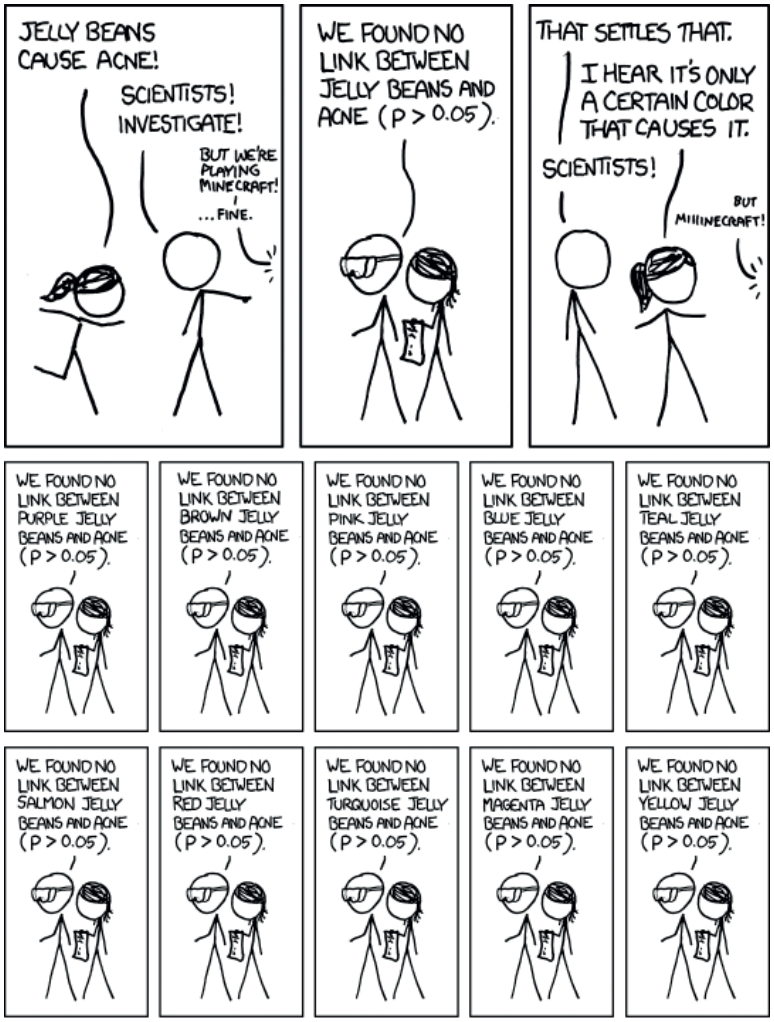
A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

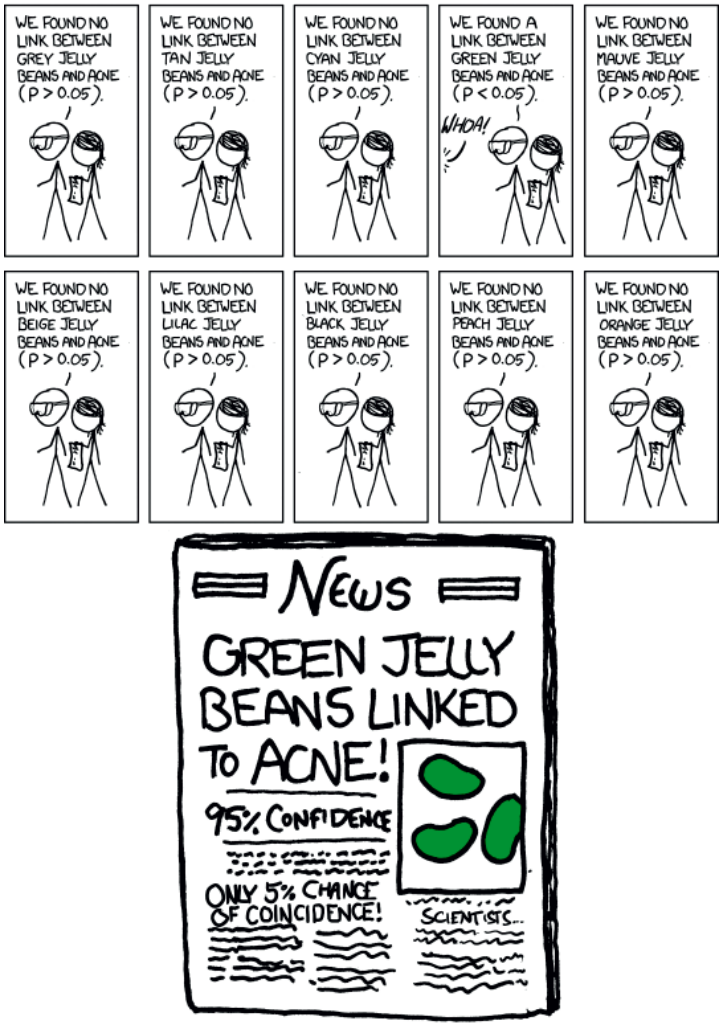
Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

Identical *t*-contrasts controlling the false discovery rate (FDR) and familywise error rate (FWER) were completed. These contrasts indicated no active voxels, even at relaxed statistical thresholds ($p = 0.25$).

Bennett, C. M., Miller, M. B., & Wolford, G. L. (2009). Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: An argument for multiple comparisons correction. *NeuroImage*, *47*(1), 125.

Multiple comparisons





http://xkcd.com/882/

The study, which Chen characterized as a "huge waste of time and money," was financed by a Johns Hopkins alumni grant to determine the effects of the compound polyphenol on cholesterol. A known antioxidant found in herbs, teas, olive oil, and wines, polyphenol was originally thought to

> In spite of the fruitless results, other researchers at Johns Hopkins expressed confidence that, in time, some positive results can be gleaned from the study. Ingels has relinquished all collected data to the university, but stressed that he will not offer any further assistance.

cholesterol. Hell, we don't know if it joins with cholesterol to form an unholy alliance to take over your gall bladder. At this point, I couldn't prove that a male donkey has nuts if they were swinging in my face."



Science Watch

result. Instead, we found out nothing."

Added Ingels: "Nothing!"

Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange an

This analysis revealed that only in ~20-25% of the projects were the relevant published data completely in line with our in-house findings

A recent report by Arrowsmith noted that the success rates for new development projects in Phase II trials have fallen from 28% to 18% in recent years, with insufficient efficacy being the most frequent reason for failure (Phase II failures: 2008–2010. *Nature Rev. Drug Discov.* **10**, 328–329 (2011))[1]. This indicates the limitations of the predictivity of disease models and also that the validity of the targets being investigated is frequently questionable, which is a crucial issue to address if success rates in clinical trials are to be improved.

Candidate drug targets in industry are derived from various sources, including in-house target identification campaigns, in-

to 'feasible/marketable', and the financial costs of pursuing a full-blown drug discovery and development programme for a particular target could ultimately be hundreds of millions of Euros. Even in the earlier stages, investments in activities such as high-throughput screening programmes are substantial, and thus the validity of published data on potential targets is crucial for companies when deciding to start novel projects.

To mitigate some of the risks of such investments ultimately being wasted, most pharmaceutical companies run in-house target validation programmes. However, validation projects that were started in our company

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must ch

Fifty-three papers were deemed 'landmark' studies … scientific findings were confirmed in only 6 (11%) cases.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability to translate cancer research to clinical success has been remarkably low[1]. Sadly, clinical

trials in onc
rate compared with other therapeutic areas.
Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will enter oncology trials. However, this low success rate is not sustainable or acceptable, and

ach to
translating discovery research into greater clinical success and impact.

Many factors are responsible for the high failure rate, notwithstanding the inherently difficult nature of this disease. Certainly, the limitations of preclinical tools such as inadequate cancer-cell-line and mouse models[2] make it difficult for even

Deming, data and observational studies

Table 1. We have found 12 papers in which claims coming from observational studies were tested in randomised clinical trials. Many of the trials are quite large. In most of the observational studies multiple claims were tested, often in factorial designs, e.g. vitamin D and calcium individually and together along with a placebo group. Note that none of the claims replicated in the direction claimed in the observational studies and that there was statistical significance in the opposite direction five times

ID no.	*Pos.*	*Neg.*	*No. of claims*	*Treatment(s)*	*Reference*
1	0	1	3	Vit E, beta-carotene	*NEJM* 1994; **330**: 1029–1035
2	0	3	4	Hormone Replacement Ther.	*JAMA* 2003; **289**: 2651–2662, 2663–2672, 2673–2684
3	0	1	2	Vit E, beta-carotene	*JNCI* 2005; **97**: 481–488
4	0	0	3	Vit E	*JAMA* 2005; **293**: 1338–1347
5	0	0	3	Low Fat	*JAMA*. 2006; **295**: 655–666
6	0	0	3	Vit D, Calcium	*NEJM* 2006; **354**: 669–683
7	0	0	2	Folic acid, Vit B6, B12	*NEJM* 2006; **354**: 2764–2772
8	0	0	2	Low Fat	*JAMA* 2007; **298**: 289–298
9	0	0	12	Vit C, Vit E, beta-carotene	*Arch Intern Med* 2007; **167**: 1610–1618
10	0	0	12	Vit C, Vit E	*JAMA* 2008; **300**: 2123–2133
11	0	0	3	Vit E, Selenium	*JAMA* 2009; **301**: 39–51
12	0	0	3	HRT + Vitamins	*JAMA* 2002; **288**: 2431–2440
Totals	0	5	52		

Goals

- Uncover bias
- Find confounders
- Detect limits, generalizability
- Validate methods/findings

- Patients heterogeneous
- Populations have different characteristics
- Sample size
- Not all data available everywhere (+ retrospective)
- Data quality (noisy, sparse, missing)

Replication doesn't validate methods

  - Need ground truth
  - Need controlled data

Today’s paper