Health Informatics

Lecture 4

Samantha Kleinberg
samantha.kleinberg@stevens.edu

- See pill ID challenge!

Uses for medical data

- Finding long-term risk factors for disease
- Drug-drug interactions, side effects (post-market)
- Population health

…

Example: Predicting cancer stage from ¨ne health forum text

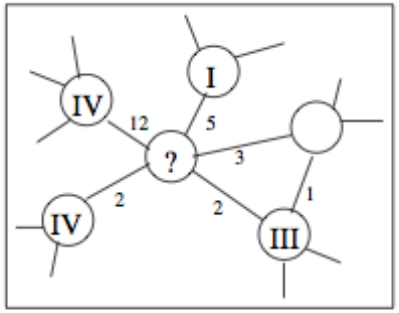


Figure 1: Nodes in the social network of forum member interaction.

Baseline				Text Based			
Stage	Precision	Recall	F	Stage	Precision	Recall	F
I	76.2	26.4	39.3	I	54.9	63.9	59.1
II	79.4	18.7	30.3	II	51.6	55.0	53.2
III	76.6	35.0	48.0	III	52.7	30.3	38.5
IV	76.4	50.7	60.9	IV	82.5	71.2	76.4
Network Based				Combined			
Stage	Precision	Recall	F	Stage	Precision	Recall	F
I	50.4	56.7	53.4	I	57.1	65.4	61.0
II	49.6	49.1	49.3	II	56.6	53.5	55.0
III	65.7	27.7	39.0	III	56.1	48.3	51.9
IV	59.3	83.7	69.4	IV	84.7	81.3	83.0

Table 3: Stage prediction results (Precision, Recall, and F-measure).

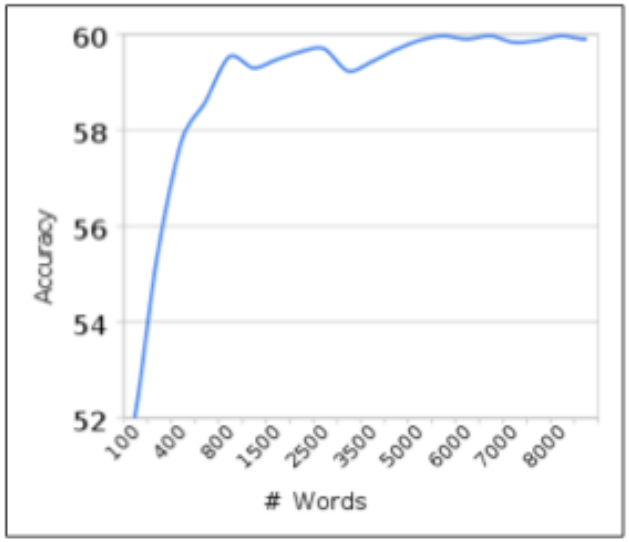


Figure 3: Overall text-based prediction accuracy against post history length.

Mukund Jha and Noemie Elhadad. Cancer stage prediction based on patient online discourse. In Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, pages 64–71, Uppsala, Sweden, July 2010

Uniqueness of Medical Data Mining

- Can't just apply methods from data mining/ ML
- Main points
  - Heterogeneity
  - Ethical, legal, social issues
  - Statistical philosophy
  - Special status of medicine

Heterogeneity of data

- Structured/unstructured, Imaging
  - Many data mining methods can handle only one of these
- How to combine qualitative/quantitative information?

Heterogeneity of data

- Importance of interpretation

- If we see CHF in record – what did clinician mean?
  - Could be suspected CHF
  - Hypothesis explaining symptoms
  - Past or family history…
- If we don't see indicator for CHF does that mean patient doesn't have it?
  - May not be billed for, may not be treated (if more pressing problems)

Heterogeneity of data

- Standardization
  - Of data (covered last week)
  - Of outcomes
- Example:
  - One group evaluates glucose control system by calculating how often glucose is within 70-150, another group uses 80-140. How to compare?

Heterogeneity of data

- Difficulty applying precise labels
- Uncertainty
  - Does a particular billing code mean patient definitely has illness?
  - When did the illness start?
- Test vs diagnosis
  - We see imperfect indicators for a disease, not the disease itself
  - For prediction, target event may be diagnosis NOT onset of disease

Ethics, legal and social factors

- MUCH more next week!

- Access to data (researchers, patients)
- Concern about liability
  - Affects what data can be collected, which tests can be done
- Privacy
  - Can we use all the data we want?
  - Can we analyze in Amazon cloud? Give text to mechanical turkers? Scrape data from password protected websites?

Statistical methods

- Do the assumptions of computational methods hold? Do we need new domain-specific ones?
  - E.g. Clinical NLP as distinct subset of NLP
- Importance of domain knowledge
- Error

Statistical methods

The data are not static

EHR systems change

Terminology changes

New tests developed

Statistical methods

Missing data

Very common and the reason for it changes interpretation

-Chronic vs acute conditions

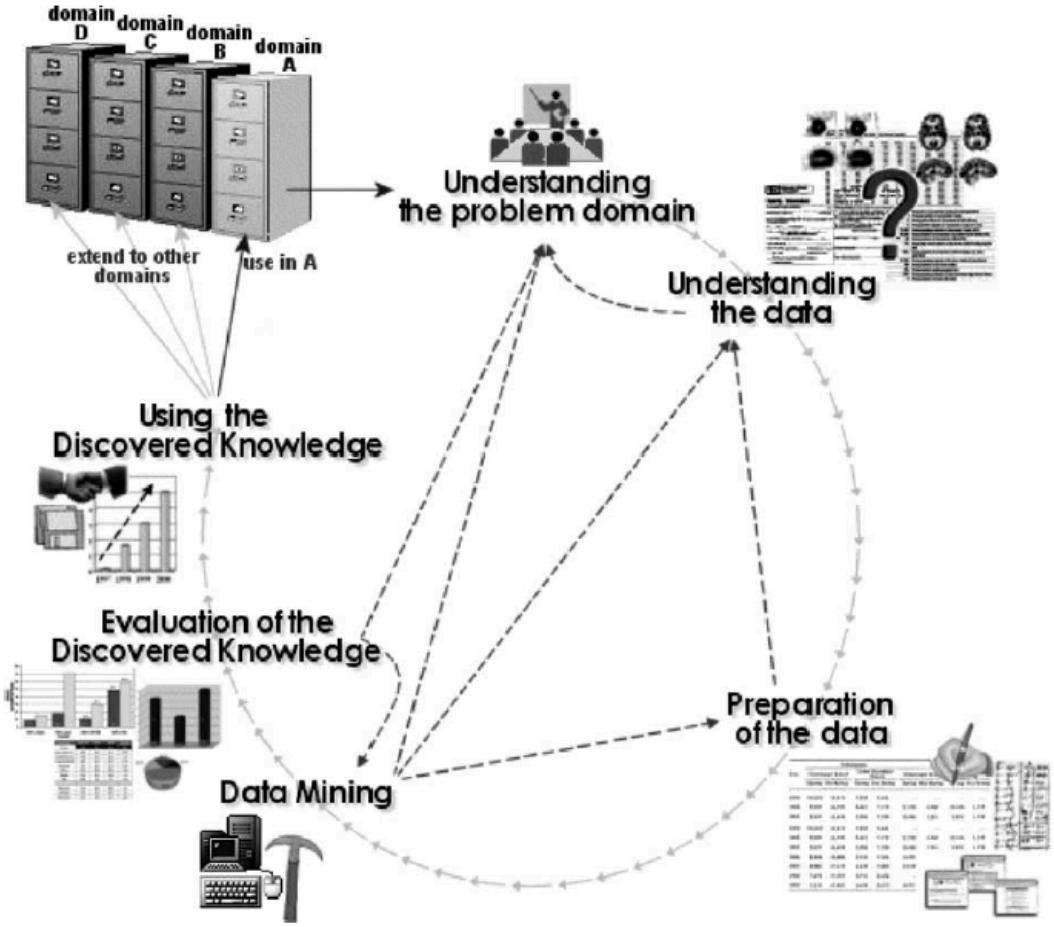
-expensive test (e.g. could be missing due to insurance coverage)

Statistical methods

Redundancy, inconsistent data

-conflicting lab results

-note says patient is Male, later says Female



Cios, K.J. & William Moore, G., 2002, Uniqueness of medical data mining, *Artificial intelligence in medicine*, 26(1-2), pp. 1-24.

Special status of medicine

- Impact/risk of false findings: good enough for a paper vs good enough to treat a patient
- Cannot do any conceivable test

More on 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

Nonstationarity

Risk Calculator for Cholesterol Appears Flawed

By GINA KOLATA
Published: November 17, 2013 | 794 Comments

Last week, the nation's leading heart organizations released a sweeping new set of guidelines for lowering cholesterol, along with an online calculator meant to help doctors assess risks and treatment options. But, in a major embarrassment to the health groups, the calculator appears to greatly overestimate risk, so much so that it could mistakenly suggest that millions more people are candidates for statin drugs.

Enlarge This Image

Mark Graham for The New York Times

Dr. Nancy Cook and Dr. Paul M. Ridker of Harvard Medical School found that a new online calculator used to assess heart treatment options

The apparent problem prompted one leading cardiologist, a past president of the American College of Cardiology, to call on Sunday for a halt to the implementation of the new guidelines.

"It's stunning," said the cardiologist, Dr. Steven Nissen, chief of cardiovascular medicine at the Cleveland Clinic. "We need a pause to further evaluate this approach before it is implemented on a widespread basis."

Other ways data may be nonstationary

- Changes in record keeping
- Changes in patient population over time
- Changes in terminology (gallop vs third heart sound; crackles vs rales)
- More accurate tests
- New diagnoses/risk factors

- Note diff between physiology and our observation of it!

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

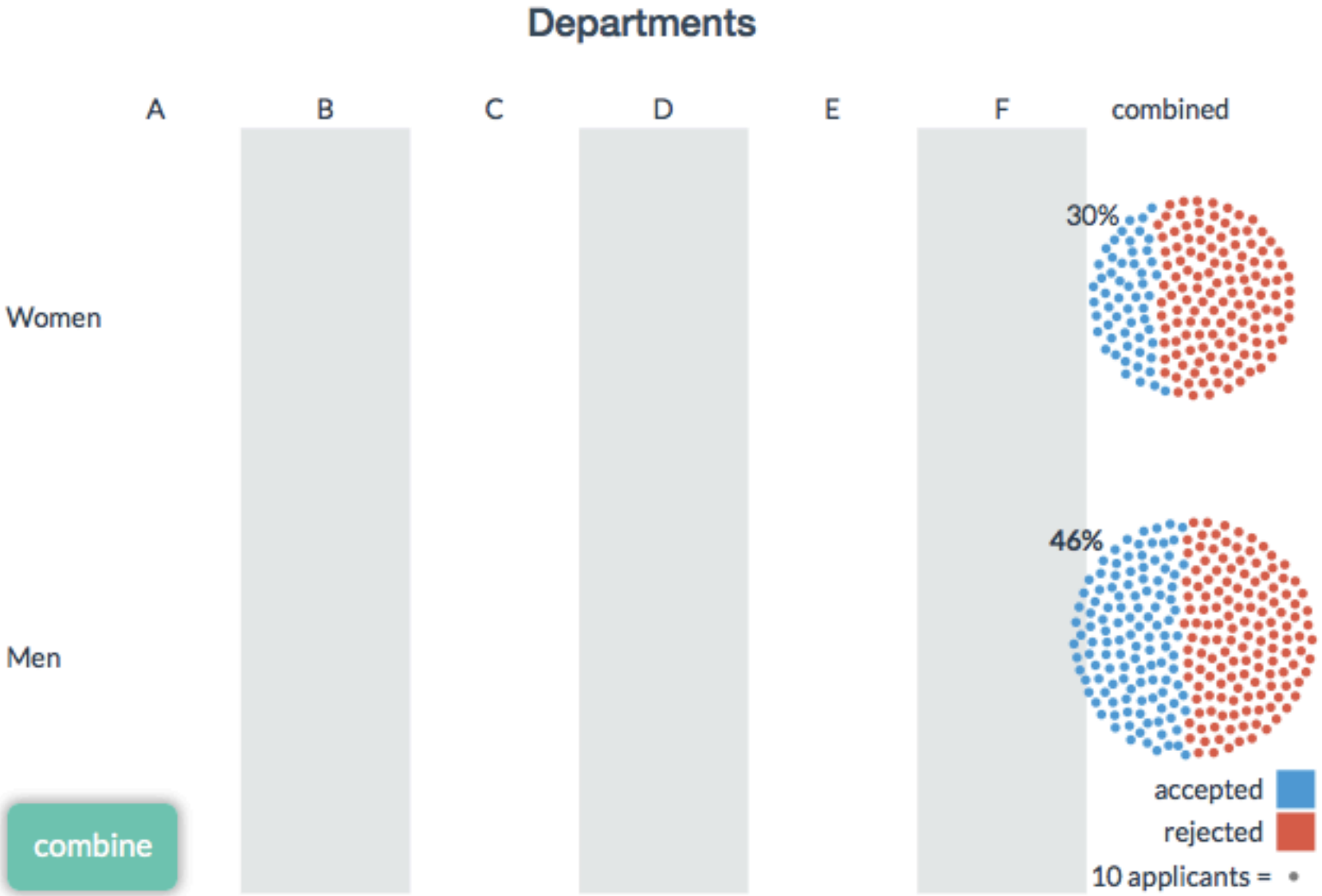
Bias in graduate admissions?

	Admit	Deny
Male	3738	4704
Female	1494	2827

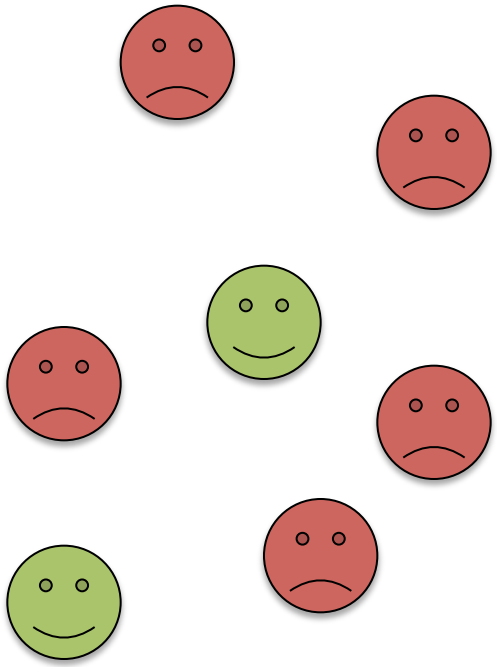
P(admitted)= 5232/12763≈0.41

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

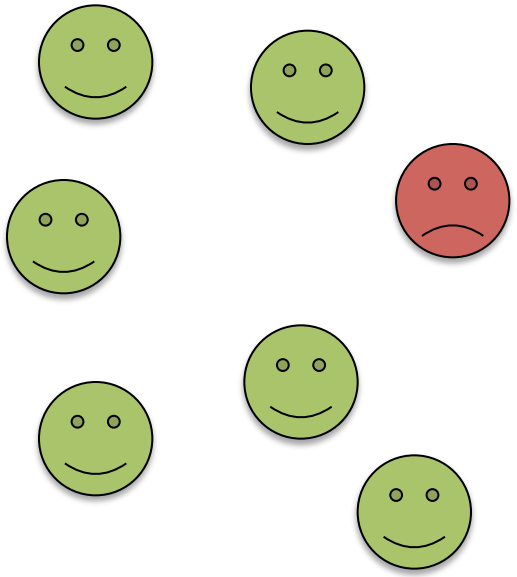
Bickel, P. J., Hammel, E. A., & O'connell, J. W. (1975). Sex bias in graduate admissions: Data from berkeley. *Science*, *187*(4175), 398



http://vudlab.com/simpsons/

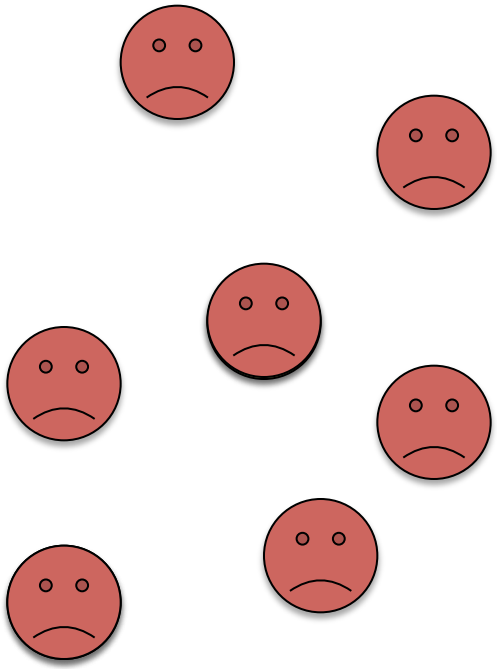


Alice's patient outcomes

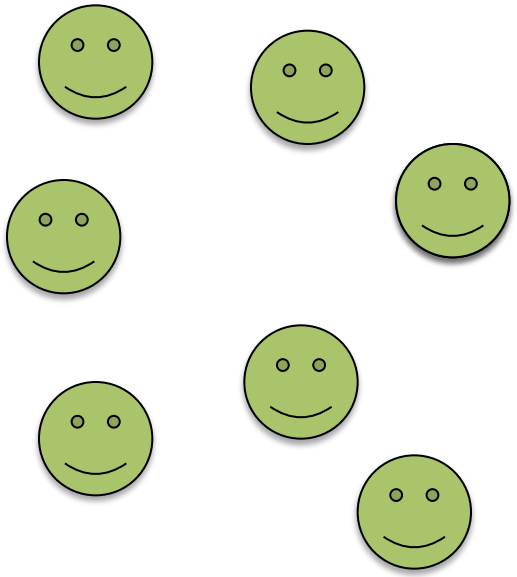


Bob's patient outcomes

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
  - Recall last week obese/morbidly obese
- 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

- 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

Is randomization always the answer?

Effects of remote, retroactive intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial

Leonard Leibovici

Department of Medicine, Beilinson Campus, Rabin Medical Center, Petah-Tiqva 49100, Israel

Leonard Leibovici *professor*

leibovic@post.tau.ac.il

BMJ 2001;323:1450–1

Abstract

Objective To determine whether remote, retroactive intercessory prayer, said for a group of patients with a bloodstream infection, has an effect on outcomes.
Design Double blind, parallel group, randomised controlled trial of a retroactive intervention.
Setting University hospital.
Subjects All 3393 adult patients whose bloodstream infection was detected at the hospital in 1990-6.
Intervention In July 2000 patients were randomised to a control group and an intervention group. A remote, retroactive intercessory prayer was said for the well being and full recovery of the intervention group.
Main outcome measures Mortality in hospital, length of stay in hospital, and duration of fever.
Results Mortality was 28.1% (475/1691) in the intervention group and 30.2% (514/1702) in the control group (P for difference = 0.4). Length of stay in hospital and duration of fever were significantly shorter in the intervention group than in the control group (P = 0.01 and P = 0.04, respectively).
Conclusions Remote, retroactive intercessory prayer said for a group is associated with a shorter stay in hospital and shorter duration of fever in patients with a bloodstream infection and should be considered for use in clinical practice.

were included in the study. Bloodstream infection was defined as a positive blood culture (not resulting from contamination) in the presence of sepsis.

In July 2000 a random number generator (Proc Uniform, SAS, Cary, NC, USA) was used to randomise the patients into two groups. A coin was tossed to designate the intervention group. A list of the first names of the patients in the intervention group was given to a person who said a short prayer for the well being and full recovery of the group as a whole. There was no sham intervention.

Three primary outcomes were compared: the number of deaths in hospital, length of stay in hospital from the day of the first positive blood culture to discharge or death, and duration of fever. Patients were defined as having fever on a specific day if one of three temperature measurements taken on that day showed a temperature of >37.5°C.

The χ^2 test was used to test for the significance of the results shown in the tables. As most of the continuous variables did not have a normal distribution, the Wilcoxon rank sum test was used for comparisons.

Results

Of 3393 patients with a bloodstream infection, 1691 patients were randomised to the intervention group and 1702 to the control group. No patients were lost to

Leibovici, L. (2001). Effects of remote, retroactive intercessory prayer on outcomes in patients with bloodstream infection: Randomised controlled trial. *Bmj*, *323*(7327), 1450-1451.

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?

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
- Interventions recorded after they're done

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

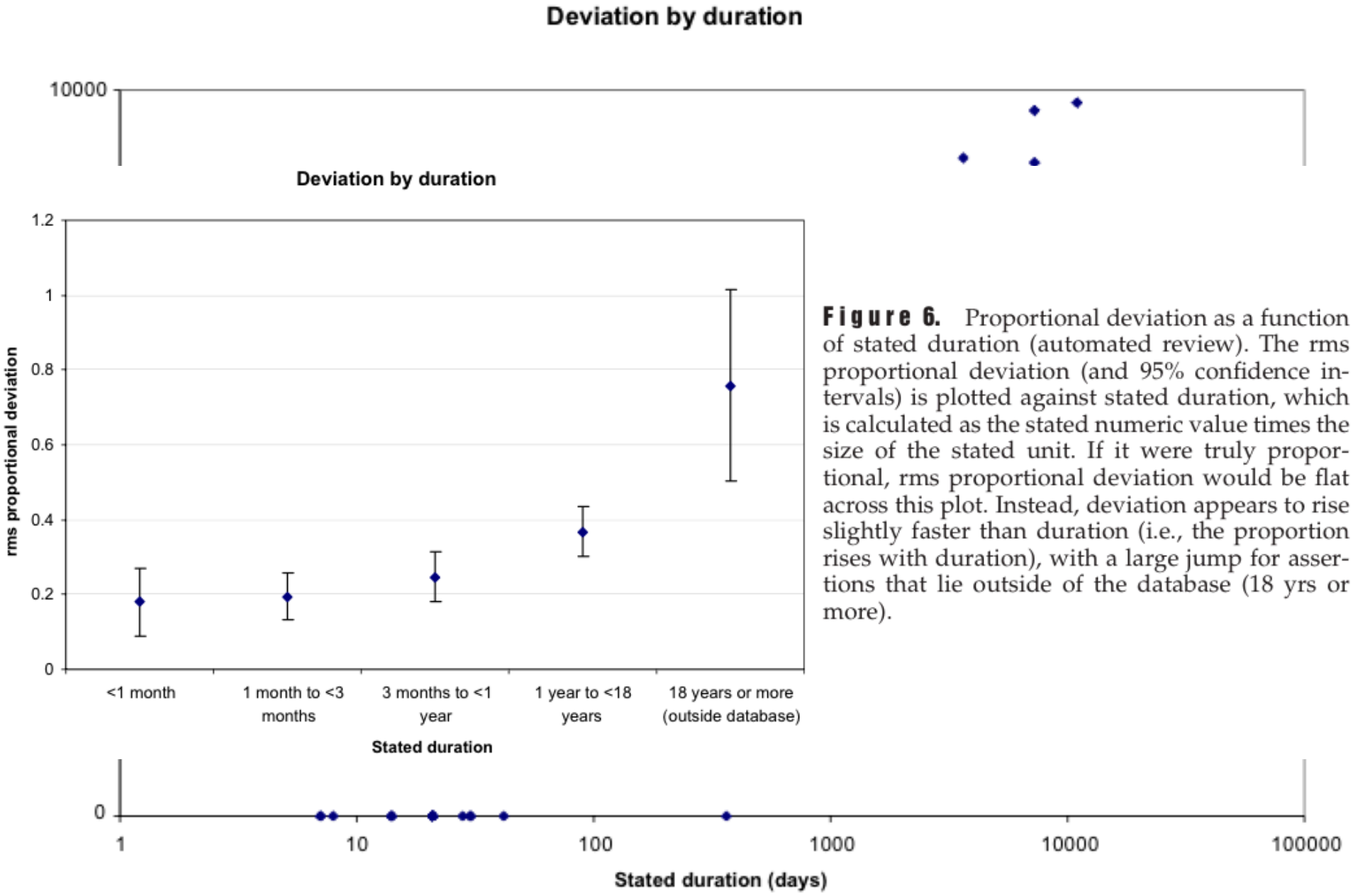


Figure 6. Proportional deviation as a function of stated duration (automated review). The rms proportional deviation (and 95% confidence intervals) is plotted against stated duration, which is calculated as the stated numeric value times the size of the stated unit. If it were truly proportional, rms 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).

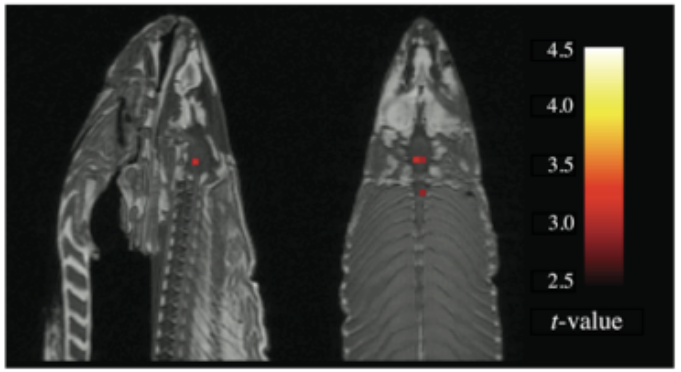
Multiple Testing

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.



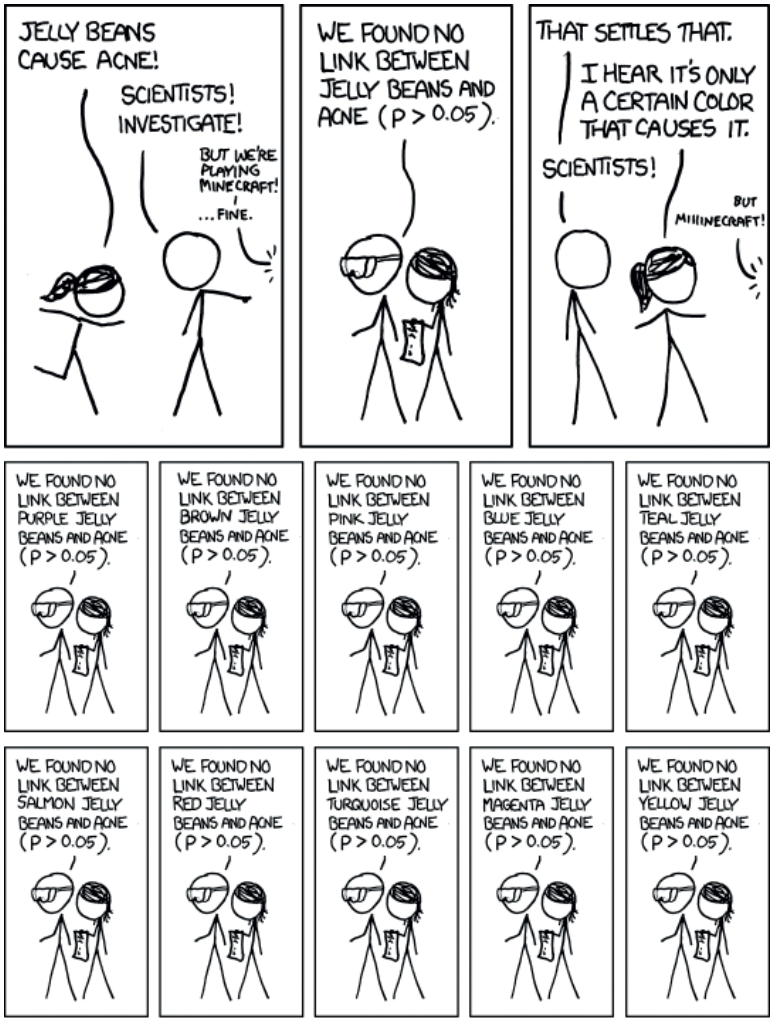
A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

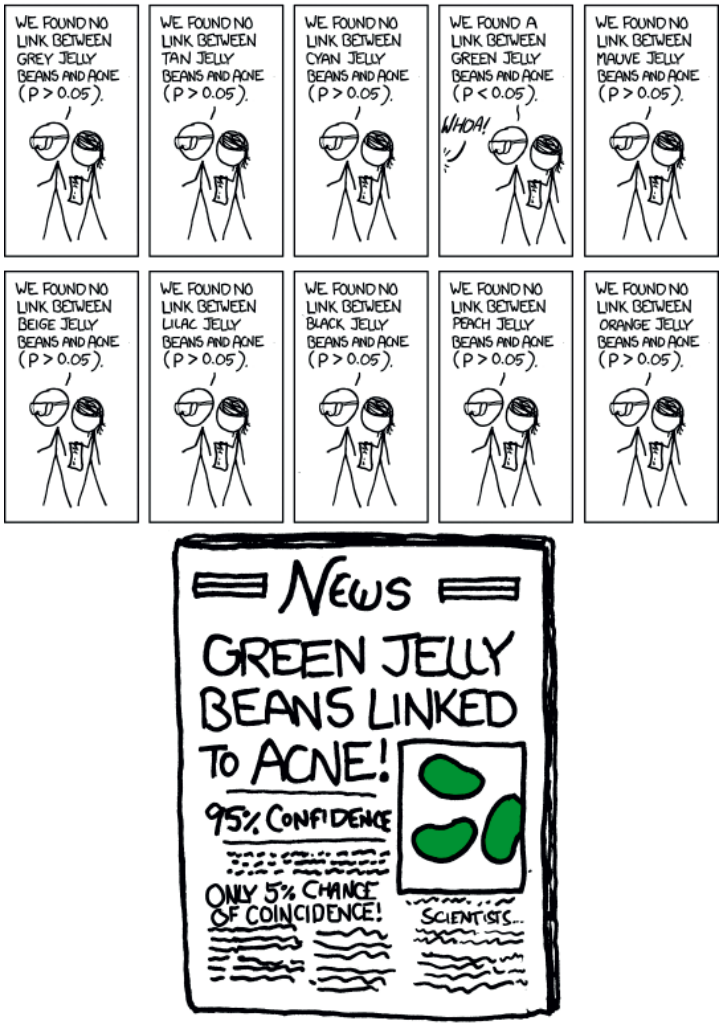
Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of 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/

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

The Government's Bad Diet Advice

By NINA TEICHOLZ FEB. 20, 2015

FOR two generations, Americans ate fewer eggs and other animal products [illegible] that [illegible] their [illegible] en

Much of the epidemiological data underpinning the government's dietary advice comes from studies run by Harvard's school of public health. In 2011, directors of the National Institute of Statistical Sciences analyzed many of Harvard's most important findings and found that they could not be reproduced in clinical trials.

[illegible] try's dietary guidelines acknowledged that they had ditched the low-fat diet. On Thursday, that committee's report was released, with an even bigger change: It lifted the longstanding caps on dietary cholesterol, saying there was

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		

Replications still somewhat rare in health informatics

Why?

Replication

Repeating same method under same conditions

Tells us method sufficiently documented and finding stable

Reproduction

Demonstrating same finding in another setting, under different conditions

Tells us about generalizability of results

Goals

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

Today's paper

- Why is reproduction of a study so difficult in biomedicine?
- What was specifically difficult in this study?
- How could we have improved the case/ control criteria or separation? (alternatively, what are some weaknesses in this?)
- Questions/criticisms?

S. Kleinberg and N. Elhadad (2013) Lessons Learned in Replicating Data-Driven Experiments in Multiple Medical Systems and Patient Populations. AMIA Annual Symposium.

- 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

For next week

T. Tsang, R. Orr, P. Lam, E. J. Comino, and M. F. Singh. Health benefits of tai chi for older patients with type 2 diabetes: The “move it for diabetes study”–a randomized controlled trial.