Causal Inference: prediction, explanation, and intervention

Lecture 11: Bias, evidence, evaluation

Samantha Kleinberg
samantha.kleinberg@stevens.edu

Administrativia

- Next week: optional final exam
- Two weeks: last class, papers due, presentations
  - 5 minute talk + 3 min q+a

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on

Cholesterol

- Previously:
  - Significant association between high HDL and reduced risk of MI
  - Increased HDL = lower risk
- Result: companies developing drugs to raise HDL

- Now: Study used natural genetic variations to test whether inherited predisposition to high/low HDL affects MI risk
  - No link between the increased HDL and lower risk, reduced HDL, higher risk

Voight, B. F. et al. (2012). Plasma HDL cholesterol and risk of myocardial infarction: A mendelian randomisation study. *The Lancet.*

SOME BACKGROUND…

Why worry about statistical significance

- Calculate ADCS, conditional probabilities, etc

- Which values are causal or noncausal?
  - Is ADCS 0.1 automatically noncausal and 0.9 definitely causal?
- Is effect in test group really bigger than in control?
  - How to compare 50% chance of death to 47%?

Choosing a threshold

- Get some numerical result
- Need to pick a cutoff point, but how?

- Flip coin 10 times
- Observe 9H 1T
- Initial hypothesis: coin is fair. How unlikely is it that we would observe 9H or something at least as extreme (10H, 10T, 9T)?

A **p-value** is the probability of getting a test statistic at least as extreme as what is observed, given that the null hypothesis is true.

Note: comparing null to alternative

Reject null for $P<\alpha$

What alpha?

- Fisher (1925) suggested using 0.05 as a threshold
  - "We shall not often be astray if we draw a conventional line at .05"

- BUT a significant result can have P>0.05, and an insignificant one <0.05



Young, S. S., & Karr, A. (2011). Deming, data and observational studies. *Significance*, *8*(3), 116-120

- Flip coin 10 times
- Observe 9H 1T
- Initial hypothesis: coin is fair. How unlikely is it that we would observe 9H or something at least as extreme (10H, 10T, 9T)?

$$P(9H1T) + P(10H) + P(10T) + P(9T1H)$$

$$P(10H) = P(10T) = (1/2)^{10} = 0.001$$

$$P(9H1T) = P(9T1H) = (1/2)^{10} \times 10 = 0.01$$

$$0.01 + 0.001 + 0.001 + 0.01 = 0.022$$

Multiple testing

- What happens if we flip 100 fair coins 10 times each?

- Should we expect to see at least one instance of 9H1T?

x=9H1T, y = events at least as extreme as x

We calculated P(y)=0.022

Now, we want P of y being true at least once in N tries

$=1-P(\text{not } y)^N$, P(not y)=1-0.022

With N=5:

$1-P(\text{not } y)^5 = 0.11$

N=50?

$1-P(\text{not } y)^{50} = 0.67$

N=100?

$1-P(\text{not } y)^{100} = 0.89$

Experiment vs. comparison significance level

- Before, rejected null for $\alpha<0.05$
- α_c = for one comparison
- α_e = for whole experiment (i.e. set of tests)

- Does $\alpha_c=0.05$ mean $\alpha_e=0.05$?

General case

- If we use α=0.05, probability of a false positive in 100 tests is:

$$1 - 0.95^{100} = 0.994$$

- Why? N tests with α_c=0.05

$$\alpha_e = 1 - (1 - \alpha_c)^N \text{ if tests independent}$$

$$\alpha_e \leq N \times \alpha_c \text{ if dependent}$$

Types of errors

	Accept null	Reject null	totals
True null H	U	V $(F+)$	m_0
False null H	T $(F\text{-})$	S	m_1
Total	$m - R$	R	m

Some terminology

- FDR V/R
- FNR T/m-R
- FWER P(V≥1)
- PCER V/m

	Accept null	Reject null	totals
True null H	U	V $(F+)$	m_0
False null H	T $(F\text{-})$	S	m_1
Total	$m-R$	R	m

Bonferroni correction

- Controls FWER P(F+≥1) but may lead to many FNs
- Recall that we can estimate

$$\alpha_e = \alpha_c \times N$$

- Thus can set

$$\alpha_c = \alpha_e / N$$

FDR control

- Compare
  - 10 tests, 2 false discoveries
  - 100 tests, 2 false discoveries
- 20% FDR vs 2% FDR
- Basic idea: focus on proportion of false discoveries

One approach: Benjamini-Hochberg

- Order m p-values so $p_1 < p_2 < \ldots < p_m$
- Where k is largest i such that

$$P(i) \leq \frac{i}{m}\alpha$$

- Reject all H_i, i=1,2…k

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289-300

Example

Say we have 15 comparisons with the following p-values:

0.0001, 0.0004, 0.0019, 0.0095, 0.0201,
0.0278, 0.0298, 0.0344, 0.0459, 0.3240,
0.4262, 0.5719, 0.6528, 0.7590, 1.000

Bonferroni: 0.05/15 = 0.0033

Example

Say we have 15 comparisons with the following p-values:

0.0001, 0.0004, 0.0019, 0.0095, 0.0201, 0.0278, 0.0298, 0.0344, 0.0459, 0.3240, 0.4262, 0.5719, 0.6528, 0.7590, 1.000

Bonferroni: 0.05/15 = 0.0033

BH: $p(4) = 0.0095 \leq \frac{4}{15} 0.05 = 0.013$

local fdr

- Result on the edge of threshold will seem less likely than it should to be a FP (the more significant results reduce the FDR)
- For a test such that FDR=0.05, that particular one has higher FDR

- Main idea: instead of using tail of distribution, assess individual results
  - What's the probability of null hypothesis conditioned on test statistic
  - Assume large number of (possibly non-independent) tests

Efron, B., Tibshirani, R., Storey, J. D., & Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, *96*(456), 1151-1160.

local fdr (continued)

- Testing how significantly individual results differ from expectations

- local false discovery rate:

  $fdr(z) = P(null|z)$

p_0 = P(null) p_1=P(non-null), $p_0 >> p_1$
$f_0(z)$ and $f_1(z)$ densities

$f(z)=p_0 f_0(z) + p_1 f_1(z)$
$fdr(z)=P(i=null|z_i=z)$
$=p_0 f_0(z)/f(z)$
$\approx f_0(z)/f(z)$

Reject null for $f_0(z)/f(z) \leq \alpha$

What's the null distribution?

- Normal
- Uniform
- Can permute data to generate

- Empirical

Why not adjust?

- Tradeoff between false discoveries and false negatives
- Arguments against adjustment (e.g. Rothman, 1990) basically center on:
  - Increased FNs are undesirable
  - Results require interpretation/follow-up

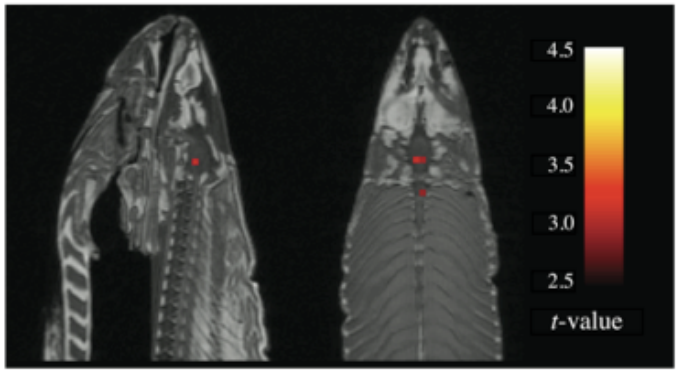
First that was the Bonferroni correction, which makes me update my belief about the results of an experiment based on how many other experiments I happen to conduct with it (and which of course implicitly assigns a low prior probability). One researcher even told me once that he has students first conduct fewer experiments so a finding has a better chance of being significant. I just walked away scratching my head. – UAI email list

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.



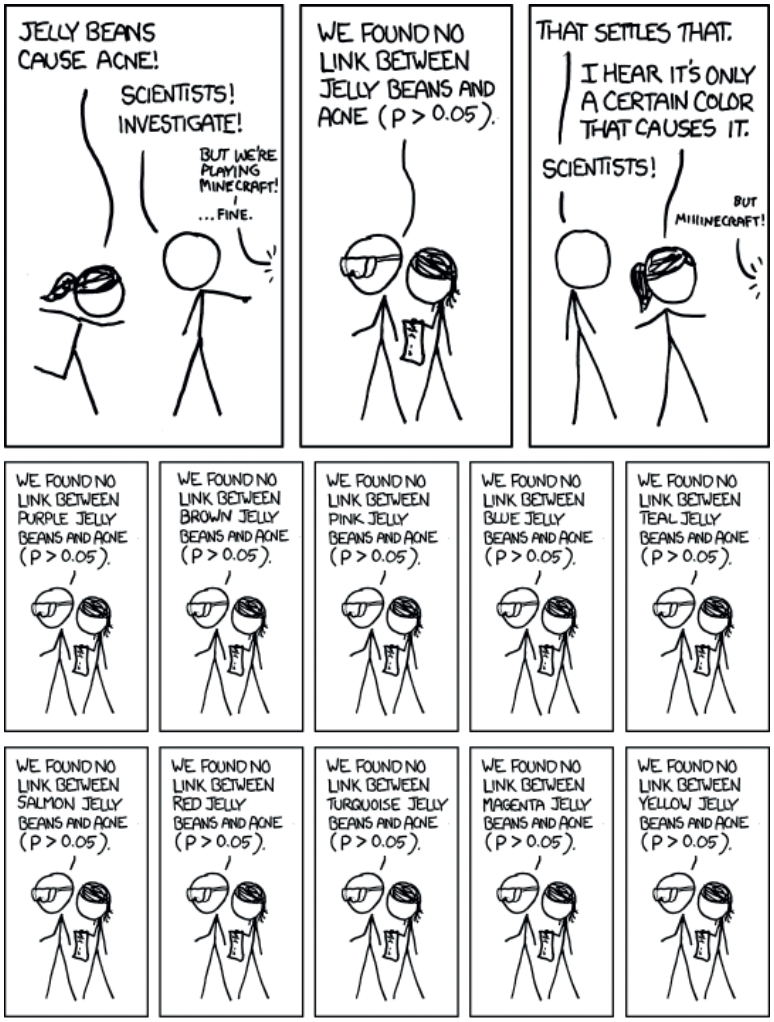
A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

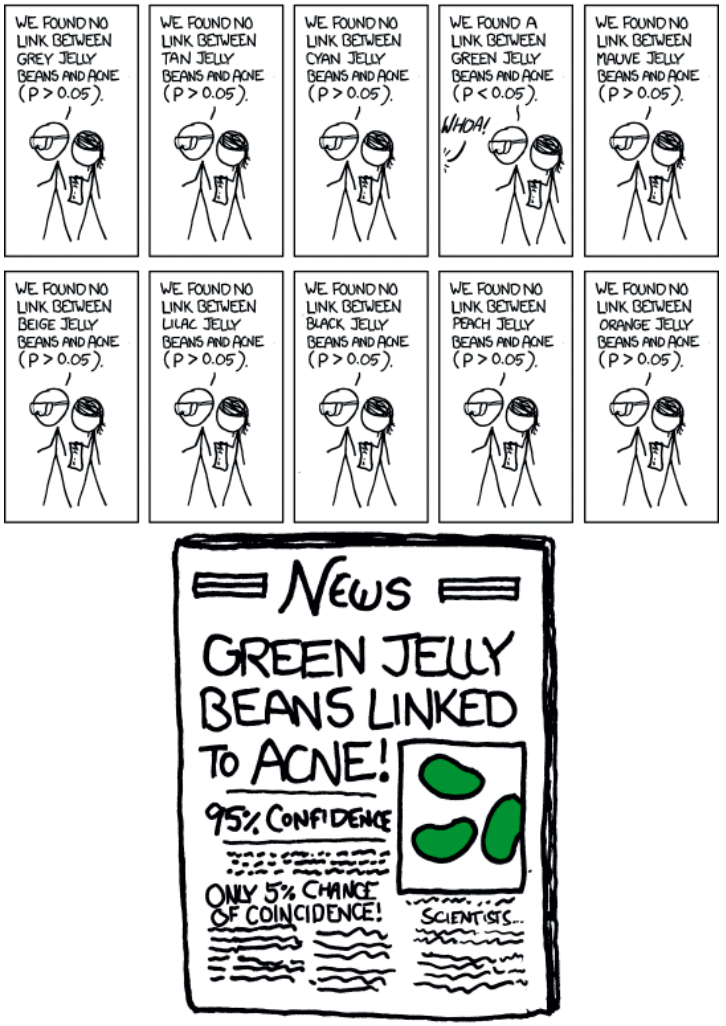
Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of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

Estimating the Reproducibility of Psychological Science

Open Science Collaboration

Abstract: Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects (Mr = .197, SD = .257) were half the magnitude of original effects (Mr = .403, SD = .188), representing a substantial decline. Ninety-seven percent of original studies had significant results ($p < .05$). Thirty-six percent of replications had significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and, if no bias in original results is assumed, combining original and replication results left 68% with significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

Citation: Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. Doi: 10.1126/science.aac4716

focus

Deming, data and observational studies

A process out of control and needing fixing

"Any claim coming from an observational study is most likely to be wrong." Startling, but true. Coffee causes pancreatic cancer. Type A personality causes heart attacks. Trans-fat is a killer. Women who eat breakfast cereal give birth to more boys. All these claims come from observational studies; yet when the studies are carefully examined, the claimed links appear to be incorrect. What is going wrong? Some have suggested that the scientific method is failing, that nature itself is playing tricks on us. But it is our way of studying nature that is broken and that urgently needs mending, say **S. Stanley Young** and **Alan Karr**; and they propose a strategy to fix it.

"There is now enough evidence to say what many have long thought: that any claim coming from an observational study is most likely to be wrong – wrong in the sense that it will not replicate if tested rigorously."

Young, S. S., & Karr, A. (2011). Deming, data and observational studies. *Significance*, *8*(3), 116-120

- Separate data cleaning and analysis
- Split sample into training/test sets
- Analysis plan based only on training set
  - Fixed prior to analysis
- Make data public

- Discuss

- Took 12 papers that analyzed observational data, where claims were also tested in RCTs
- No results were replicated in the RCT
  - Observation= 52 positive claims
  - RCT = 5 negative claims, 47 no significance
- Why?

Bias

- Recall selection criteria for RCT, methods for randomization
- Reason for medication vs its effects

Multiple modeling

Box 3. Bisphenol A

The US Center for Disease Control assayed the urine of around 1000 people for 275 chemicals, one of which was bisphenol A (BPA). One resulting claim was that BPA is associated with cardiovascular diagnoses, diabetes, and abnormal liver enzyme concentrations. BPA is a chemical much in the news and under attack from people fearful of chemicals. The people who had their urine assayed for chemicals also gave a self-reported health status for 32 medical outcomes. For each person, ten demographic variables (such as ethnicity, education, and income) were also collected. There are 275 × 32 = 8800 potential endpoints for analysis. Using simple linear regression for covariate adjustment, there are approximately 1000 potential models, including or not including each demographic variable. Altogether the search space is about 9 million models and endpoints[11]. The authors remain convinced that their claim is valid.

Young, S. S., & Karr, A. (2011). Deming, data and observational studies. *Significance*, *8*(3), 116-120.

Hill's viewpoints

The Environment and Disease: Association or Causation?

by Sir Austin Bradford Hill CBE DSC FRCP(hon) FRS
(*Professor Emeritus of Medical Statistics, University of London*)

Amongst the objects of this newly-founded Section of Occupational Medicine are firstly 'to provide a means, not readily afforded elsewhere, whereby physicians and surgeons with a special knowledge of the relationship between sickness and injury and conditions of work may discuss their problems, not only with each other, but also with colleagues in other fields, by holding joint meetings with other Sections of the Society'; and, secondly, 'to make available information about the physical, chemical and psychological hazards of occupation, and in particular about those that are rare or not easily recognized'.

Nine features to consider when evaluating causality in cases where experimentation is not possible

Mix of: showing cause makes a difference, and that there's a potential mechanism

Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, *58*, 295-300

(1) Strength

- How strong is the association?

- Intuitively, if presence of adverse event in people taking a drug is nearly the same as that in the general population, this is less convincing than if there is a significant increase

(1) Strength continued

- Weak causes
  - Secondhand smoke and lung cancer
  - Unidentified subgroups
  - Rare events
- Strong non-causes
  - Down syndrome and birth order

(2) Consistency

- Are results replicable?
- Is relationship observed by different groups, using different methods?

- Ex: running and side stitch

(2) Consistency continued

- Inconsistent causes
  - Unidentified features necessary for cause to be effective
  - Mosquito bites and malaria
  - Inconsistency in cause ≠ inconsistent study results
- Consistent non-causes
  - Common study flaw

(3) Specificity

- Does cause lead to one effect or a group of effects?
- Specificity also in magnitude

- Examples
  - smoking causing illness vs. smoking causing lung cancer
  - One SNP causing many phenotypes

(3) Specificity continued

- Not necessarily 1 to 1 relationship
  - Factors can increase risk of multiple disease
  - Drugs have many side effects

(4) Temporality

- Does the cause precede the effect?
- Is the delay consistent with how cause might work?

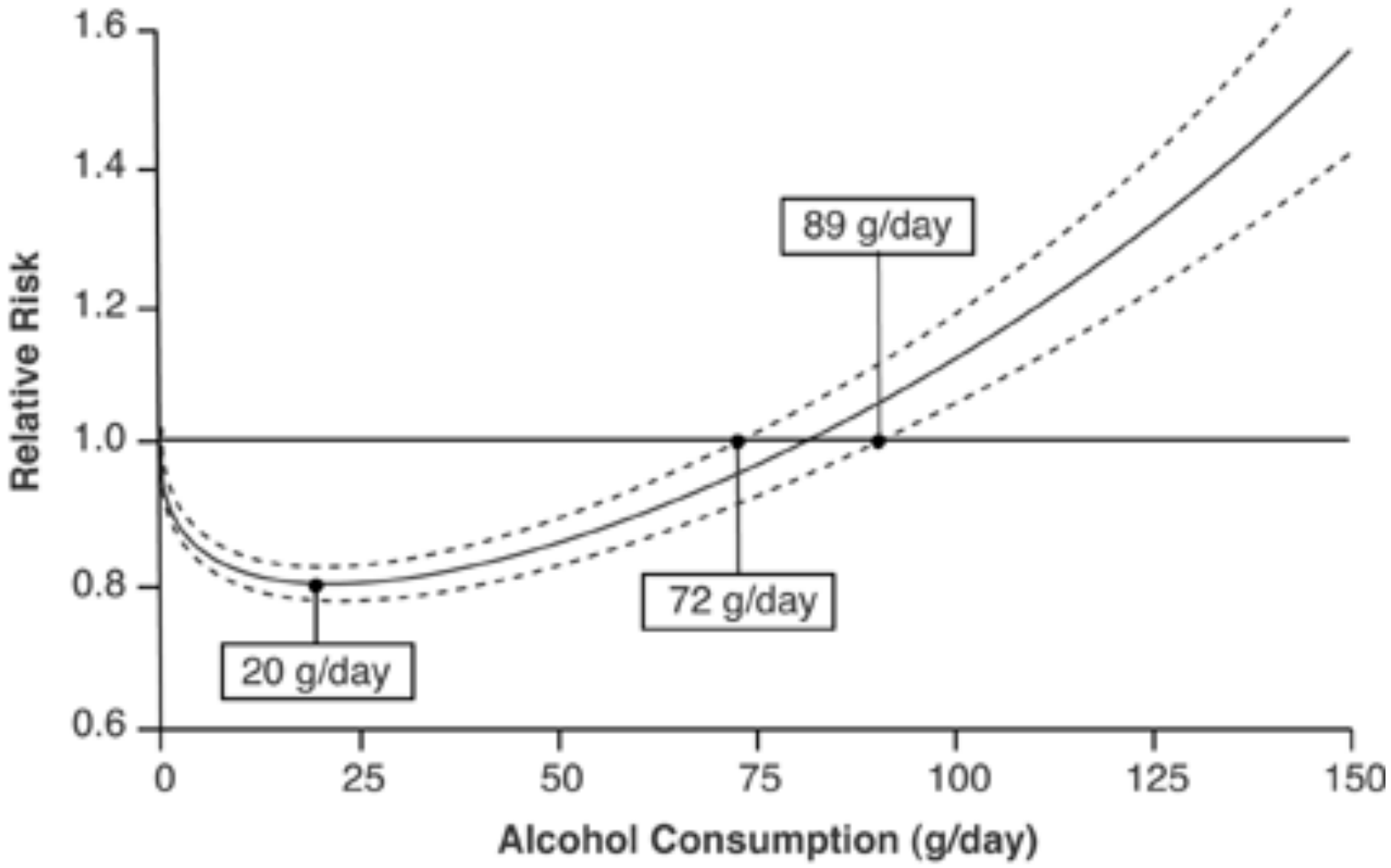
- Ex: Occupation and illness

(5) Biological gradient

- Does level of effect or risk of it increase with increase in level of cause?
- Dose response curve
- As more people in an area are exposed to risk factor, does disease incidence increase?

- Study involving 121,342 people
- Found each daily increase of 3oz of meat associated with 12% increased risk of cv death a 10% increase in cancer death

J-shaped curve



Rehm, J., Gmel, G., Sempos, C. T., & Trevisan, M. (2003). Alcohol-related morbidity and mortality. *Mouth*, *140*(208), C00-C97

(6) Plausibility

- Is there a potential mechanism that could connect the cause and effect given what we currently know?

- Depends on current knowledge

(7) Coherence

- Does the relationship conflict with known facts?

- Note that current knowledge may be wrong

Plausbility vs coherence

- Plausibility: we can conceive of a way the relationship could work given what we know
- Coherence: the relationship does not conflict with what we know

(8) Experiment

- Are there experimental results supporting the relationship?

- If we intervene by increasing/introducing the cause, does the effect result?

- Ex: does lung cancer rate change as smoking rate does?

(8) Experiment continued

- Human vs. animal (+ representativeness)
- Randomization, blinding
- Sample size, population

(9) Analogy

- Example: Know that HPV causes cervical cancer, so may be more likely to accept virus as cause of another type of cancer
- Example 2: If we see effect of listing calorie content on orders at restaurants, more plausible that listing salt content can change behavior too

Here then are nine different viewpoints from all of which we should study association before we cry causation. What I do not believe – and this has been suggested – is that we can usefully lay down some hard-and-fast rules of evidence that *must* be obeyed before we accept cause and effect. None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a *sine qua non*. What they can do, with greater or less strength, is to help us to make up our minds on the fundamental question – is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect?

Hill notes

- NOT a checklist
- NOT features of relationship itself

- Methods for evaluating associations, recognizing causes.
- Still need to think about quality of evidence of each type
  - Recall last week: variations in RCTs

Hill + causal inference

1. Strength Probability e.g. ADCS, Granger
2. Consistency Probability, regularity
3. Specificity Probability
4. Temporality Temporal priority
5. Biological gradient Mill's methods
6. Plausibility Mechanisms
7. Coherence Mechanisms
8. Experiment Interventions, RCTs
9. Analogy Mechanisms

Hill's Viewpoints applied to smoking

1. Strength – Rate of lung cancer (and death from LC) in smokers much higher than in non-smokers, 9-10x higher.
2. Consistency – Same link found by many studies (retrospective and prospective), in many populations (men & women, geographic)
3. Specificity – Death rate generally inc. among smokers, but death from LC inc. substantially more
4. Temporality – People usually smoke before developing LC/dying
5. Biological gradient – Death rate/LC incidence increases linearly with increased smoking
6. Plausibility – Theory of tissue damage eventually leading to cancer
7. Coherence – Fit with what was known at time
8. Experiment – Tar on the ears of lab rats produced cancer, showing carcinogens are present
9. Analogy – Animal → human

For next week

Final exam!!