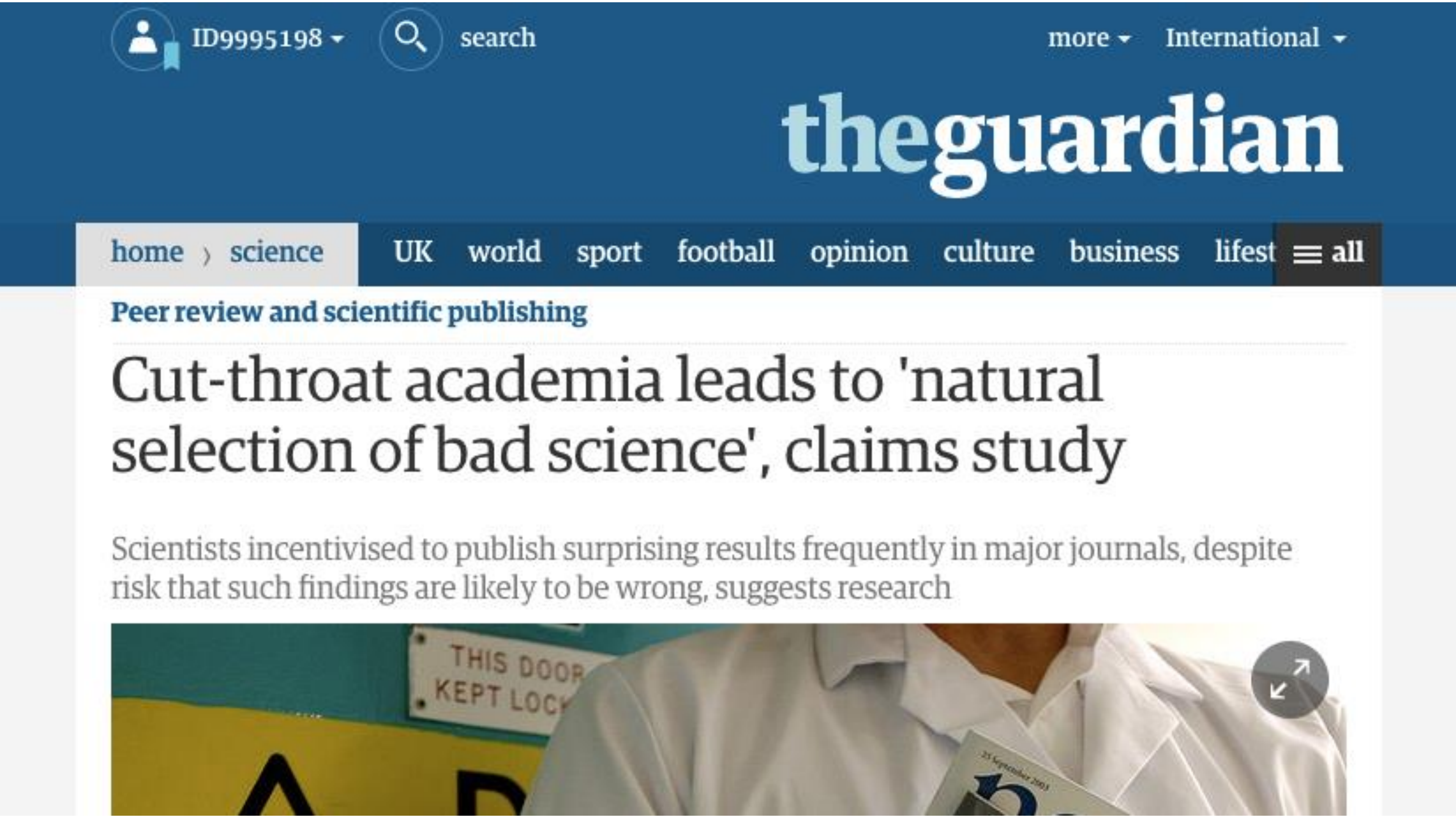
Hypothesis testing

or

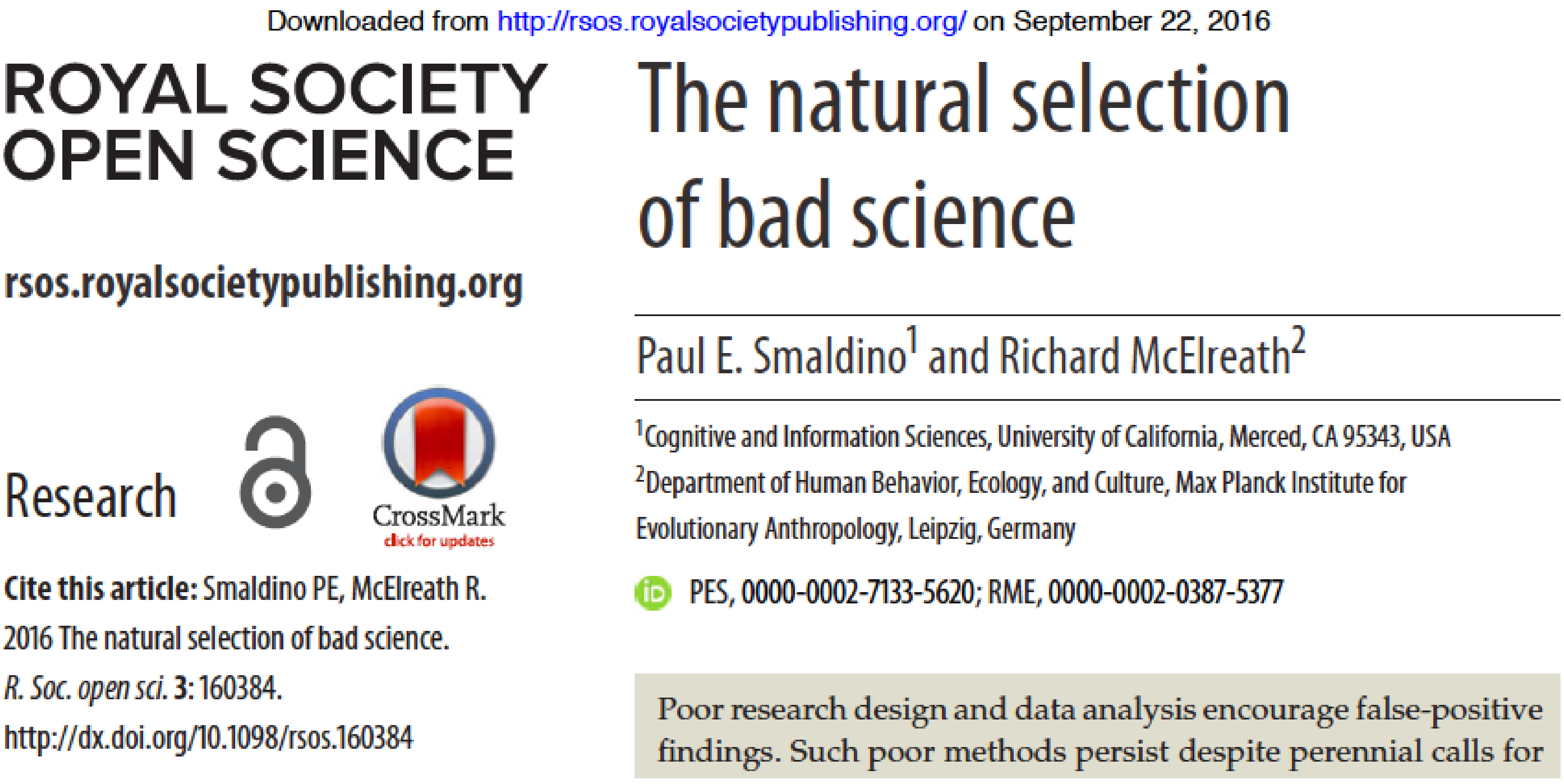
How to interpret a P-value



Peer review and scientific publishing

Cut-throat academia leads to 'natural selection of bad science', claims study

Scientists incentivised to publish surprising results frequently in major journals, despite risk that such findings are likely to be wrong, suggests research



ROYAL SOCIETY
OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

CrossMark
click for updates

Cite this article: Smaldino PE, McElreath R. 2016 The natural selection of bad science. *R. Soc. open sci.* **3**: 160384. http://dx.doi.org/10.1098/rsos.160384

The natural selection of bad science

Paul E. Smaldino[1] and Richard McElreath[2]

[1]Cognitive and Information Sciences, University of California, Merced, CA 95343, USA
[2]Department of Human Behavior, Ecology, and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

iD PES, 0000-0002-7133-5620; RME, 0000-0002-0387-5377

Poor research design and data analysis encourage false-positive findings. Such poor methods persist despite perennial calls for

Lack of understanding of how to interpret the results of a hypothesis test is a key problem

Objectives

At the end of this session you should

- Understand role of chance in analytic epidemiology
- Understand how to interpret a P-value in context of hypothesis testing

Aim of analytic epidemiology

- To determine the relationship between exposure and outcome
- Contrast between..
- ..testing
  - (is there an association)
- ..and, estimation
  - (what is the size of the effect)

Why does it matter?

- When association is detected possible causes are:
  - Chance
  - Bias
  - Confounding
  - Real
- Considerable discussion in literature about causes of false positives
  - the role of bias and confounding usually discussed in depth
  - but role of chance rarely discussed appropriately
- Major problem in dietary, molecular, genetic and clinical epidemiology (and quantitative science in general)

Anatomy of a statistical test

- Define null hypothesis: H_0
  - H_1 is the alternative hypothesis
  - H_0 and H_1 are two sides of the same coin
- Define α
  - Significance level at which null hypothesis will be rejected
- Collect data
- Calculate appropriate test statistic
  - Compare observed data with data expected if H_o were true
- Obtain P-value
- Accept or reject null hypothesis
  - convention is to reject if $P<0.05$
- Most "significant" results will be false positive results using $P=0.05$ threshold

Note that

prob null is true + prob alternative is true = 1

The test statistic

- Many different statistical tests that produce a test statistic from different statistical distributions
  - z, t, chisq, F
- The test statistic has the same fundamental format

$$\frac{\text{some measure of observed data}}{\text{some measure of expected data if null were true}}$$

- Number with no units
- Because the underlying distribution is known can obtain the probability of getting a test statistic that is as large or larger than the one obtained

Definition of P-value

- Probability ..
- ..of data as or more extreme than those observed..
- ..occurring IF null hypothesis is true
- Pr(observed data|null hypothesis)

The P-value is a conditional probability

Example 1

- H_o there is no association between smoking and lung cancer
- Carry out a perfect study with no bias/confounding
- Do statistical test
- Look up P-value
- P = .03
- Reject null hypothesis

- Is this a true positive or a false positive?

Example 2

- H_o women are shorter than men on average
- Carry out a perfect study with no bias/confounding
- Do statistical test
- Look up P-value
- P = .03
- Reject null hypothesis
- Accept the alternative that women are taller than men

- Is this a true positive or a false positive?

Is the P-value what we really want to know?

P-value is the probability of data observed (or more extreme data) IF the null hypothesis is true

But, what we really want to know is..

.. the probability that the null hypothesis is true given the data observed..

..or, probability that the alternative hypothesis (H_1) is true given the data

What we want to know is the probability that women are shorter than men given the (perfect) data we have collected

What we have calculated from the data is the probability of obtaining these data if women are shorter than men (i.e. P-value)

A bit of maths

P-value = Pr(data | null)

What we want to know = Pr(null | data)

Pr(null | data) ≠ Pr(data | null)

- Interpretation of P-value as probability that the null hypothesis is correct is thus erroneous
- Understanding this is key to interpretation of the P-value

$$\underbrace{\Pr(\text{null} \mid \text{data})}_{\text{What we want to know}} = \frac{\overbrace{\Pr(\text{data} \mid \text{null})}^{\text{P-value}}.\overbrace{\Pr(\text{null})}^{\text{Prior probability of null}}}{\underbrace{\Pr(\text{data})}_{\text{1- statistical power}}}$$

But…….

- We don’t know whether or not the null hypothesis is true
- And so the P-value cannot be interpreted in a useful way
- Frequentist statistics need to be interpreted in a Bayesian way

How not to interpret a P-value

"If the new drug has a positive effect, frequentist statistics indicate how likely it is that the effect occurred by chance, the *P* value. If the probability that chance was to blame is less than, say, 5% (a *P* value of 0.05), the researcher has a positive answer, a certain degree of confidence, and possibly a new, effective drug."

Beckman M. 2006 JNCI 98;1512-13

Pharoah P. 2007 JNCI 99:332-3

Outcomes of a statistical test

- H_1 true and H_o rejected (a)
  - Correct decision
- H_o true and H_o rejected (b)
  - Incorrect decision
  - Type I error
- H_1 true and H_o accepted (c)
  - Incorrect decision
  - Type II error
- H_o true and H_o accepted (d)
  - Correct decision

	Truth		
Test	H_1 true	H_0 true	
H_0 rejected	a	b	(a+b)
H_0 accepted	c	d	(c+d)
	(a+c)	(b+d)	N

	Truth		
Test	H_1 true	H_0 true	
H_o rejected	a	b	(a+b)
H_o accepted	c	d	(c+d)
	(a+c)	(b+d)	N

Probabilities

- Probability of type I error

  = b / (b+d)
- Probability of a type II error

  = β = c / (a+c)
- Statistical power

  = 1 - β = a / (a+c)

Statistical power

- Probability of rejecting H_0 when H_1 is true
- Depends on effect size and α
- Can be calculated

	Truth		
Test	H_1 true	H_0 true	
H_0 rejected	a	b	(a+b)
H_0 accepted	c	d	(c+d)
	(a+c)	(b+d)	N

What we want to know

- The probability that H_0 is true given that H_0 rejected

  = b / (a+b)
- Or, the probability that H_1 is true given H_0 rejected

  = a / (a+b)
- But, we do not know the marginal totals and therefore cannot calculate

The false positive reporting probability

- The probability that H_0 is true given a statistically significant finding
- Described by
- Wacholder and colleagues 2004 *JNCI*: 96; 432-34
- Called the false positive reporting probability (FPRP)

What’s needed to estimate FPRP?

- If we know the prior probability that H_1 is true, and
- Power of the statistical test
- Can estimate FPRP, because
- Prior = (a+c)/N
- Power = a / (a+c)

Estimating FPRP

	H_0 false	H_0 true	
H_0 rejected	80	45	125
H_0 accepted	20	855	875
	100	900	1000

Prior = 0.1

Power = 0.8

Type I error = 0.05

FPRP = 45 / 125 = 36%

Demonstration of FPRP spreadsheet

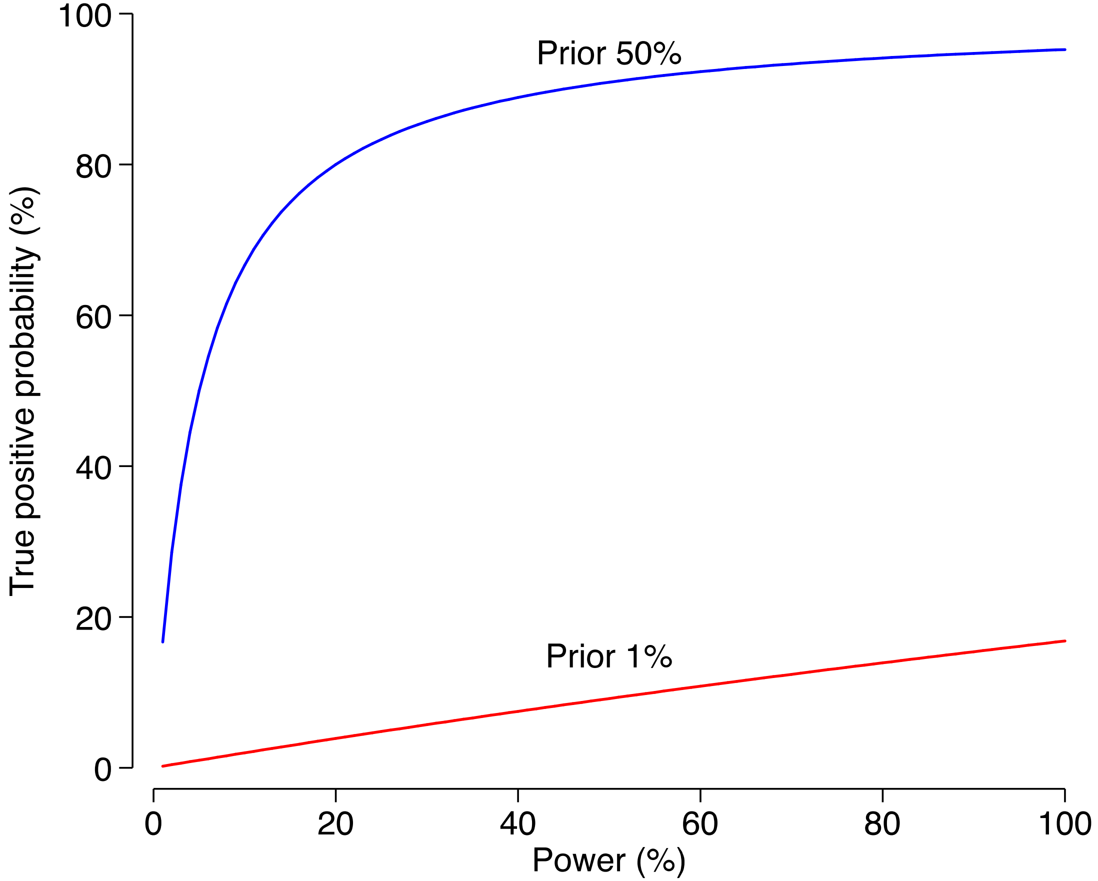
Problems

- Don't know what the prior is
- Has to be guestimated
- But orders of magnitude may be reasonable
- Prior known in context of a clinical trial
  - if there is true equipoise the prior must be 50%
- For many exposures (e.g. molecualr epi, dietary epi, genetic associations) priors are very low

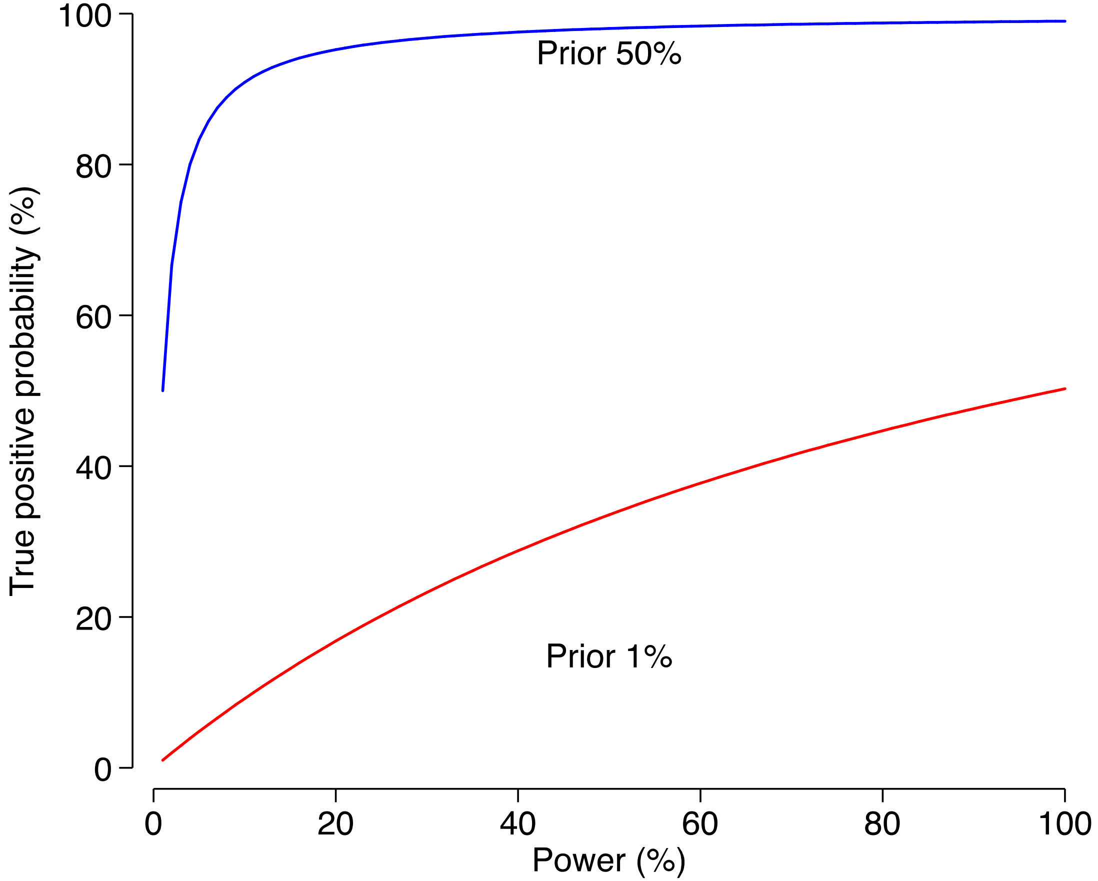
Prior probability that H_1 true

- Unknown, and may be virtually impossible to estimate
- Factors to consider
  - Biological plausibility
  - Prior data on same or similar hypothesis
- Investigators often have unrealistic priors, particularly when they have strong biological hypothesis
- Lung cancer and smoking prior v high
- Lung cancer and star sign prior very low (?0)

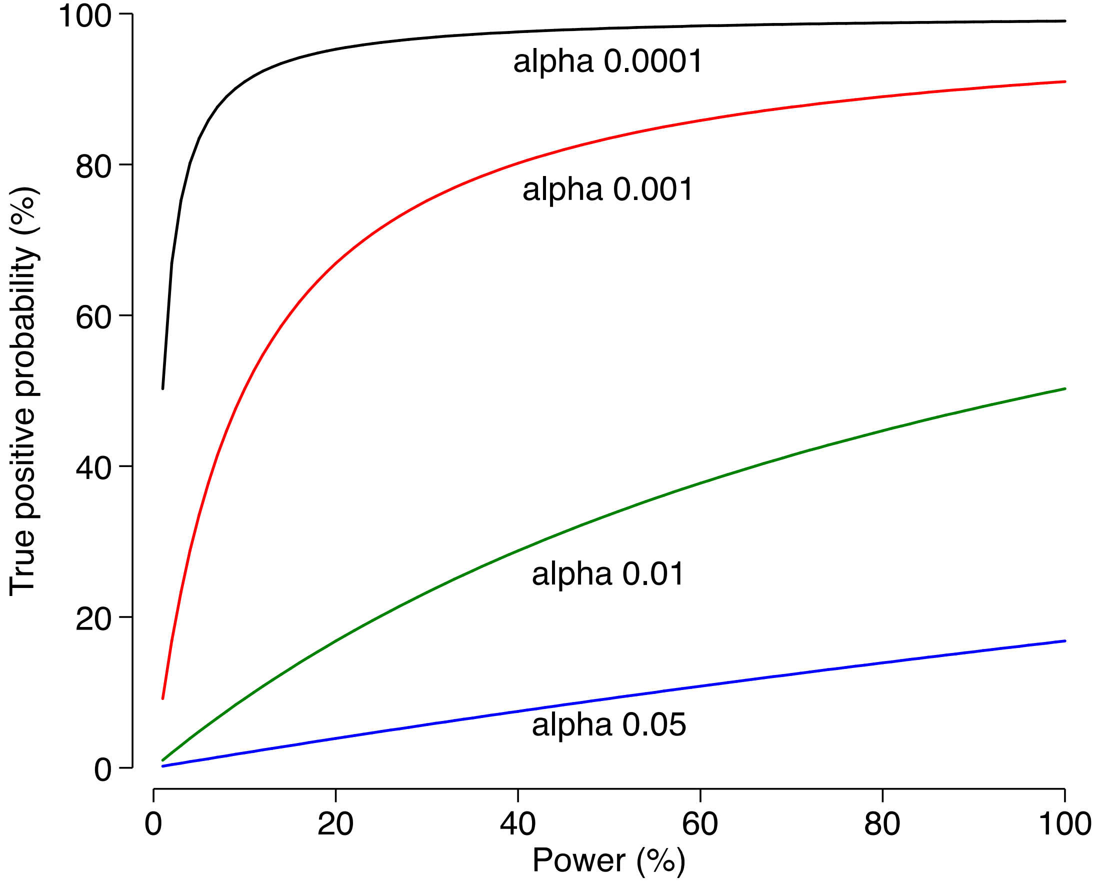
True positive probability for type I error 0.05



True positive probability for type I error 0.01

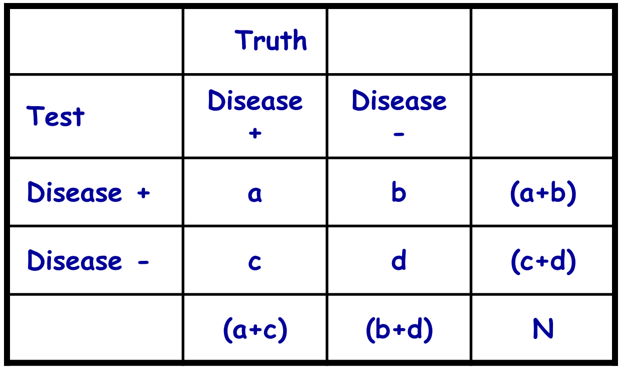


True positive probability for prior 1%



False positive reporting probability

- The probability that H_0 is true given a statistically significant finding
- Depends on
  - Prior probability that H_1 true
  - Power of test to detect an effect
  - Both of these depend on effect size
- A Bayesian interpretation of a frequentist statistic



	Truth		
Test	Disease +	Disease -	
Disease +	a	b	(a+b)
Disease -	c	d	(c+d)
	(a+c)	(b+d)	N

Parallels with a clinical test

- Test sensitivity = a / (a+c) ~ power
- Test specificity = d / (b+d) ~ 1-Type I error

What we really want to know is probability that someone who test positive has disease

  – Positive predictive value

- PPV = a / (a+b)
- Depends on disease prevalence = (a+c) / N ~ prior

Key points

- A small study generating a given P-value (declared significant) is more likely to be a false positive than a large study generating the same P-value

  Power of the study is important

- The prior probability that the alternative hypothesis is correct is usually small (often very small)

  Under these circumstances, most associations declared significant at P<0.05 will be false positives

- Threshold of 0.05 is totally inadequate for most epidemiology

Questions?