

# Using Bayes to separate evidential wheat from chaff

**Robert Matthews**  
**Aston University, Birmingham UK**

**email: [rajm@physics.org](mailto:rajm@physics.org)**

# Declaration of Interest

Robert Matthews declares he has no conflicts of interest and has received no funding for what follows

# A crisis of confidence in evidence

ESSAY ⓘ

## Why Most Published Research Findings Are False

Article

Metrics

Related Content

Comments: 26

John P. A. Ioannidis

 To add

**Publication Bias in Reports of Animal Stroke Studies Leads to Major Overstatement of Efficacy**

# The nature of the problem

- Ioannidis (*JAMA* 2005)
  - High-impact (>1000 citation) studies, 1990-2003
  - 45 studies claimed “significant” effects ( $p < 0.05$ )
  - 14 (31%) failed to replicate in direction/magnitude

# Observational studies

- Ioannidis (2005)
  - 6 observational studies making “significant” claims
  - 5 (83%) failed to replicate in magnitude/direction
- Young & Karr (*Significance* 2011)
  - Observational studies later tested via RCTs
  - 12 studies, 52 observational claims
  - 52 (100%) of claims failed to replicate

# What's going wrong ?

- Poor design (eg inadequate sample size)
- Data-dredging of large samples
- Bias (recruitment, recall, publication...)
- Confounding
- Blunders/Fraud
- Something else....?

Something wrong with concept of “significance” ?

# Problems with significance testing

- P-values widely used but...
  - They don't mean what they seem to
  - They assume the null, and so can't also confirm/refute it.
  - Commonly misinterpreted: " $P < 0.05$ " does NOT mean:
    - " $P(\text{null}) < 0.05$ "
    - " $P(\text{effect is real}) > 0.95$ "

# Why this matters

- Davey Smith & Ebrahim (*BMJ* 2002):
  - Paired 133 measures from a study against each other
  - Over 8,800 pairs → should give ~ 88 “significant” relationships at  $p < 0.01$  level;
  - Actually gave 3,000+ (!)
  - *Vast majority had zero plausibility*

# Confidence intervals are better...

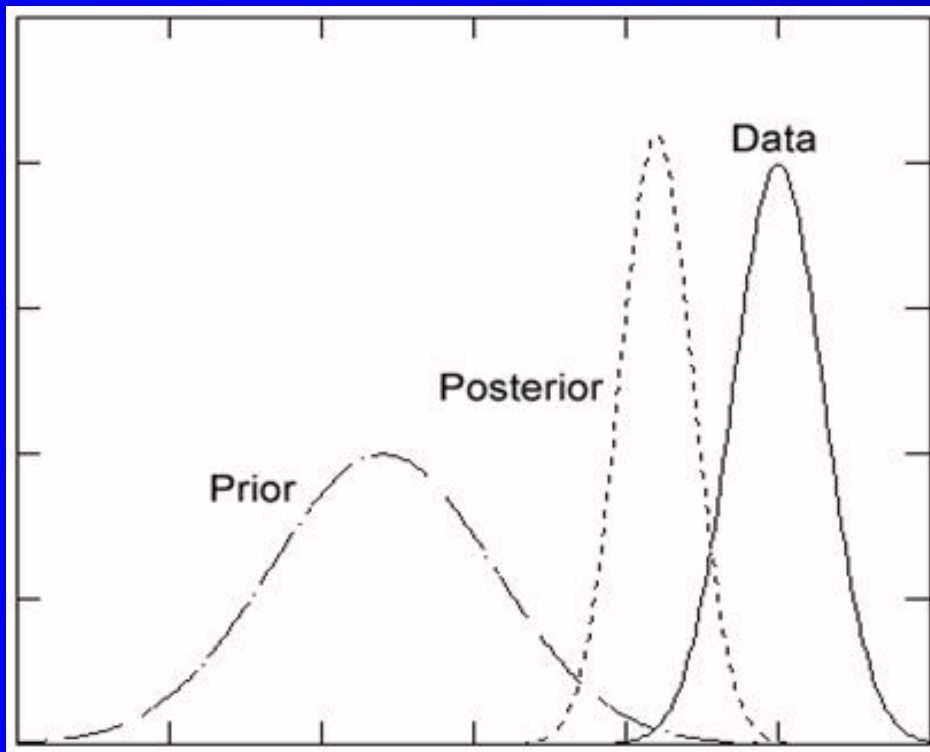
- 95% CIs are (somewhat) clearer: in repeated sampling, 95% of them include parameter
- They are more informative:
  - $N = 100$ :  $OR = 0.3$ ,  $p \sim 0.02$  95% CI = (0.13, 0.90)
  - $N = 700$ :  $OR = 0.6$ ,  $p \sim 0.02$ , 95% CI = (0.41, 0.90)
- But width of CI often ignored
- Still unclear how to include plausibility.

# So: what to do ?

- Bayesian methods
  - Rigorous framework for including prior knowledge
  - Results “mean what they seem to”
  - Allows plausibility to be assessed *quantitatively*

# Bayesian methods

Prior evidence + data  $\Rightarrow$  updated evidence



Prior:  $C_{Pr} (L_{Pr}, U_{Pr})$

Data:  $C_D (L_D, U_D)$

**Bayes: Prior + data**

Post:  $C_{Pt} (L_{Pt}, U_{Pt})$

# Using existing insight via Bayes

## Anistreplase and heart attacks

- GREAT study (*BMJ* 1992): OR = 0.47 (0.23, 0.97)  
**Significant - but *plausible* ?**
- Pocock & Spiegelhalter (*BMJ* 1992): “OR ~ (0.6, 1.0)”
  - Prior evidence + data (via Bayes):  
**OR ~ 0.73 (0.6, 1.0)**
- Morrison *et al.* (*JAMA* 2000): OR = 0.83 (0.70, 0.98)

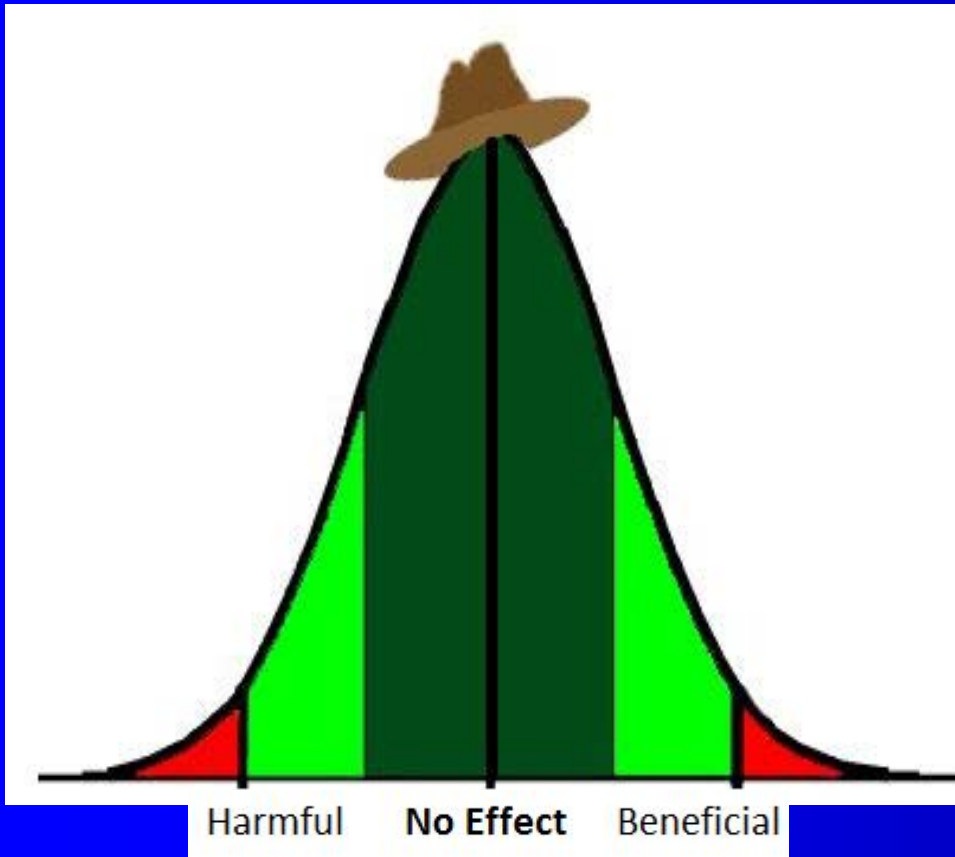
# Plausibility analysis via Bayes

- Challenges:
  - Whose prior should be used ?
  - “Optimism bias” of experts (Chalmers & Matthews *Lancet* 2006)
  - What if no previous studies (“Out of the blue”)

# Plausibility analysis via Bayes

- Use Bayes to ask:
  - *“What would extant evidence need to show, for new result to be plausible ?”*
- Calculate prior needed for result to be plausible at 95% level
- Then ask: *“Is this supported by extant evidence?”*
- Bayes  $\Rightarrow$  Need location/shape of prior distribution

# The “Open-Minded Skeptic” prior



- “Open-minded”: symmetric (ORs  $\leq$ / $\geq$  1 given equal weight)
- “Skeptic”: centred on No Effect (i.e. OR = 1)
- A Normal guy, so unassuming and easy to work with....

# Plausibility analysis via Bayes

- Input – data: 95% CI bounds for OR:  $L_D$ ,  $U_D$
- Output - Credible Prior Value (CPV): ORs needed to convince the Open-Minded Skeptic.
- Online calculator - [bit.ly/PlausibilityTest](http://bit.ly/PlausibilityTest) (courtesy Dr John C. Pezzullo, [www.statpages.org](http://www.statpages.org))

## Plausibility Analysis rules

- If OR = 1 lies outside quoted 95% CI, result is *statistically significant*
- If ORs at least as impressive as CPV are plausible, result is also *credible at 95% level*

# Plausibility analysis via Bayes

- **“Out of the blue” findings**
  - Study should not demand prior belief in an even more impressive effect.
  - $\Rightarrow$  Study's central value must exceed the study's CPV
  - If it does not, study lacks evidential weight to make its case.

# Fibre and cancer: wheat or chaff?

- Schatzkin *et al.* 2007: cereal fibre “modestly protective”: 95% CI = 0.86 (0.76, 0.98)

## *Plausibility Analysis*

- Is it *statistically significant* ? **Yes**:  $U_D < 1.0$
- Is it *credible* ? Only if CPVs  $< 0.93$  ; yes, and consistent with central value.
- This “significant” finding is also credible
- Meta-analysis (*Aune et al* BMJ 2011) confirmed credibility

# Fibre and cancer: wheat or chaff?

- Nomura *et al.* 2007: *fruit* fibre also protective: 95% CI: 0.88 (0.78, 0.99)

## *Plausibility Analysis*

- Is it *statistically significant* ? **Yes**:  $U_D < 1.0$
- Is it *credible* ? Only if CPVs  $< 0.75$  are already plausible. Not so, and inconsistent with central value
- Treat this “significant” finding as “weak”.
- Meta-analysis (Aune *et al* 2011): no association

# Beyond mere “significance”

- The current approach of assessing evidence is a simple dichotomy:

<b>95% CI excludes 1.0 or <math>p &lt; 0.05</math></b>	<b>95% CI includes 1.0 or <math>p &gt; 0.05</math></b>
Significant evidence; Important	Not significant evidence; Unimportant

# Gauging evidence via CPVs

Strength of evidence	Credible Prior Value (CPV)
<b>Category A</b> (“Confirmatory”)	Significant <u>and</u> CPV in line with previous studies of hypothesis
<b>Category B</b> (“Credible”)	Significant <u>and</u> CPV in line with results expected for hypothesis; central value consistent with CPV
<b>Category C</b> (“Preliminary”)	Significant, <u>but</u> CPV lacking support from other studies; central value consistent with CPV
<b>Category D</b> (“Weak”)	Significant, <u>but</u> CPV lacks support from other studies; central value NOT consistent with CPV

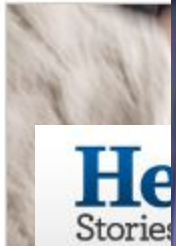
# One new threat to health...



Sugar-free diet drinks

# Made a lot of headlines...

Diet soft drinks may raise stroke risk, study says



News Sp

Breaking News

Last Updated:

Diet so



Risk?  
Diet Soda

Font size:  
A A A

Share this:

in, MD



Study



Print



RSS

...le drinking soda regularly

# Diet sodas: a threat to health?

- Gardener *et al.* 2011: vascular event risk increased to 1.48 (1.03, 2.12)

## *Plausibility Analysis*

- Is it *statistically significant* ? **Yes**:  $L_D > 1.0$
- Is it *credible* ? Only if CPV > 2.4 is already plausible. No evidence AND central value inconsistent with CPV.
- This “significant” finding is Category D evidence: “Weak”

# Separating wheat from chaff

- Use of Bayes moves us on from pass/fail dichotomy (significance *and* plausibility assessed)
- Makes the most of information in 95% CIs
- Compels sceptics *and* advocates to be transparent and quantitative in their views.
- Reduces unreliability of findings in journals (and media)

*“Editors are people who separate the wheat  
from the chaff – and print the chaff ”*

Adlai Stevenson

Thank you

# Appendix

## The mathematical basis of Credible Prior Values (CPVs)

The underlying theory for the calculation of CPVs can be found in the appendix to Matthews (2001b, see below), where it is used to derive so-called Credible Prior Intervals (CPIs) for use in assessing the plausibility of results from Randomized Controlled Trials (RCTs). The ILSI presentation described the concept of the Open-minded Skeptical Prior, which allows the same approach to be applied more widely, notably to the findings of observational studies, and to “out of the blue” results (Matthews [*in prep*]).

Matthews (2001b) includes a nomograph which allows both CPIs and CPVs to be estimated from quoted 95% CIs. An online calculator has also been created by J Pezzullo at [Statpages.org](http://Statpages.org), and is available at [bit.ly/PlausibilityTest](http://bit.ly/PlausibilityTest). While this uses ORs, RRs can be substituted if the absolute event probability is low.

# Selected references

1. Aune, D. *et al* (2011) *Dietary fibre, whole grains, and risk of colorectal cancer: systematic review and dose-response meta-analysis of prospective studies* Br Med J 343 d6617
2. Gardener, H. *et al* (2011) *Soda Consumption and risk of vascular events in the Northern Manhattan Study*. Int Stroke Conf. Los Angeles, CA. Abstract P55.
3. Matthews, R A J (2001a) *Why should physicians care about Bayesian methods ?* J Stat Plan Inf 94 (1) 43
4. Matthews, R A J (2001b) *Methods for assessing the credibility of clinical trial outcomes* Drug Inf J 35 (4) 1469 (Available online at: [bit.ly/CredibilityAnalysisPaper](http://bit.ly/CredibilityAnalysisPaper) )
5. Spiegelhalter, D *et al.* (2004) *Bayesian Approaches to Clinical Trials and Healthcare Evaluation* (Wiley) Ch 3 p75 *et seq*
6. Young, S; Karr, A (2011) *Deming, data and observational studies* Significance 8 (3) 116.