The Analysis of Credibility

A framework for moving researchers beyond NHST

Robert Matthews

Dept of Mathematics, Aston University

rajm@physics.org

NISS Webinar 19 November 2019

Where we are after 80+ years of complaining

Despite our protestations, researchers typically…

- Are sticking with p-values and/or 95% CIs
- Still think output of NHST is easy to interpret
- Want *every* study to have simple output ("H_1 true/false")

A pragmatic way forward

- Accept the above, but make p-values/95% CIs
  - **More** informative
  - **Less** prone to misinterpretation
  - **More** nuanced in their implications

A Bayesian approach

- Answers relevant inferential questions
- Clear incorporation of weight of evidence
- Extracts more insight from data summaries
- Sets new findings in context of prior insight

A Bayesian approach

- Answers relevant inferential questions
- Clear incorporation of weight of evidence
- Extracts more insight from data summaries
- Sets new findings in context of prior insight

Analysis of Credibility (AnCred)

(Matthews 2018, 2019)

Standard Bayesian approach

Prior $\otimes$ *f*(Evidence) ➔ Posterior

Jack Good (1950)

"What prior would give a credible posterior ?"

1. Posterior $\otimes$ f^{-1}*(Evidence)* ➔ Prior
2. Assess this prior in context of existing knowledge

“Fair-minded challenge” of claims

- **Input**: 95% CI summary statistic of finding
- **Analysis:** subject evidence to *fair-minded challenge:*

Significant result: challenge by fair-minded **sceptic** of H_1

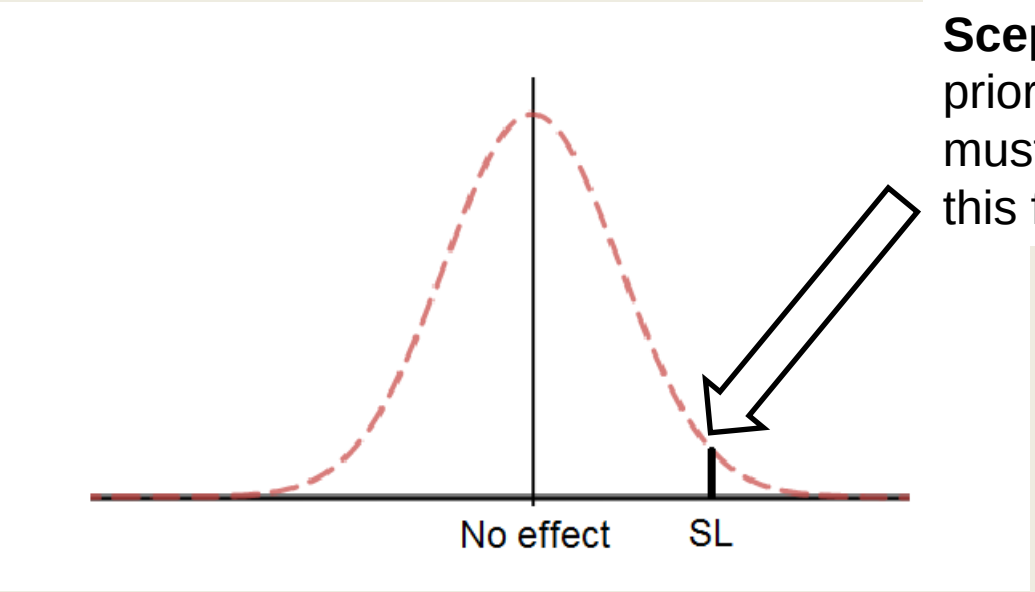
*What level of scepticism would make result **not credible** ?*

Non-significant result: challenge by fair-minded **advocate**:

*What level of advocacy would make result **credible**?*

Statistically significant results

The Fair-minded Sceptic

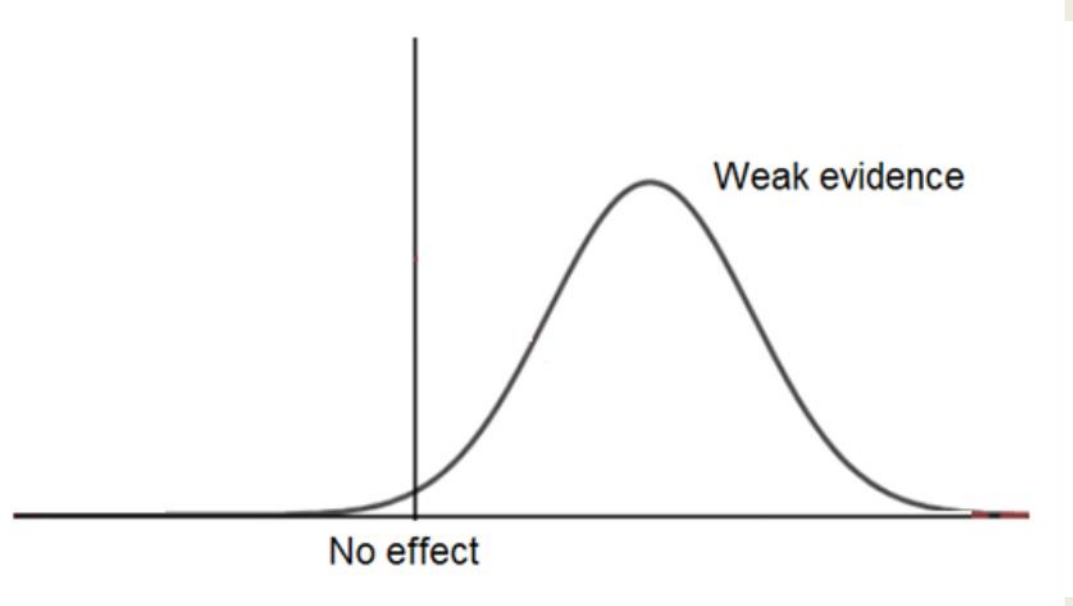


Scepticism Limit: prior evidence must lie **beyond** this for credibility

- Centred on no effect (“Sceptic”)
- 95% tails set by strength of evidence are equipoise (“Fair-minded”)

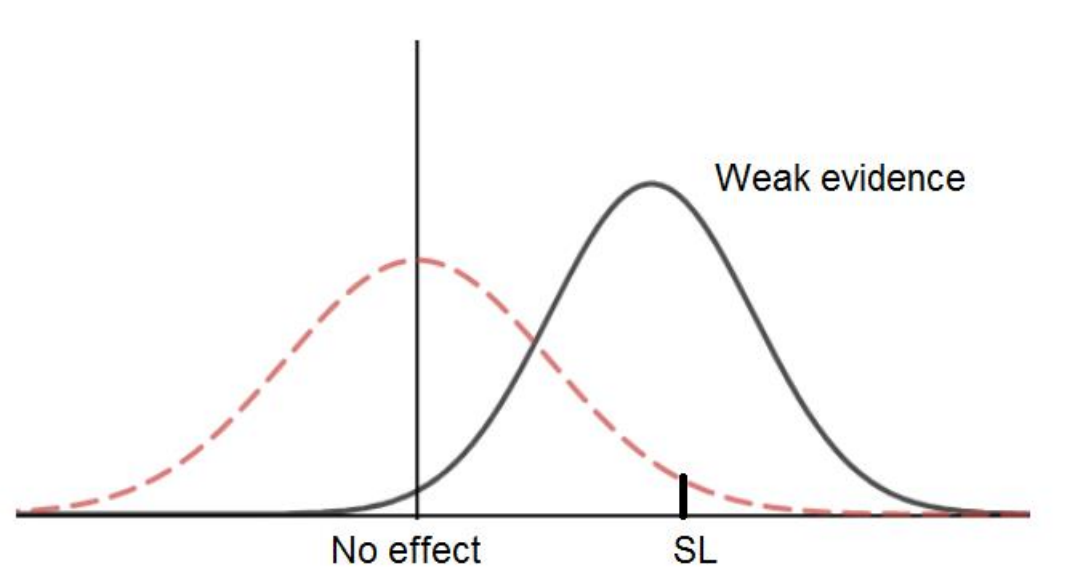
Statistically significant results

Example: weak evidence



Statistically significant results

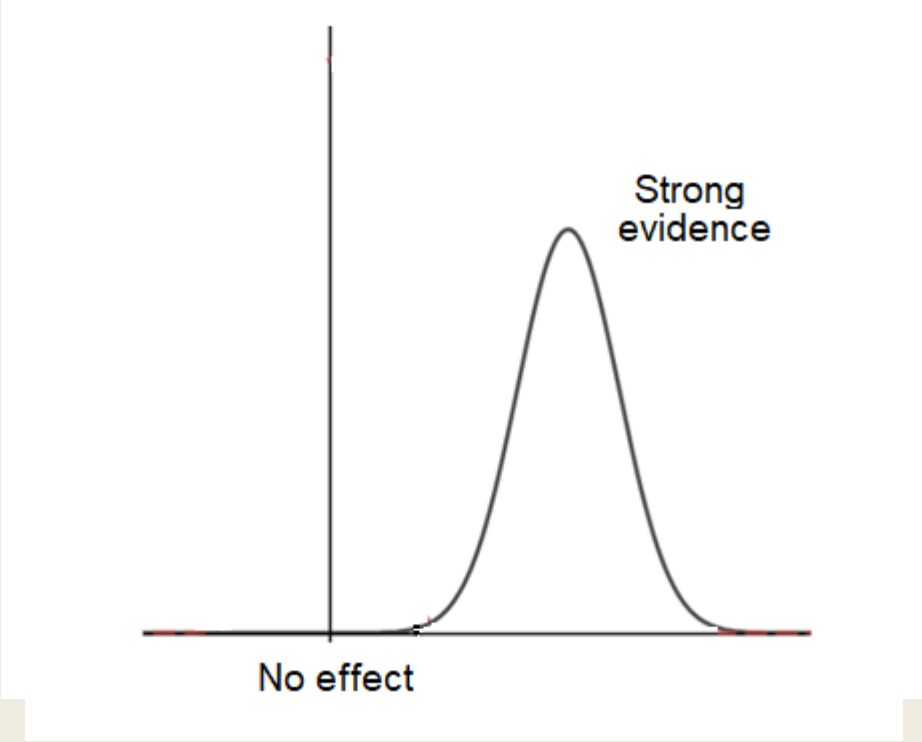
Sceptic's response to evidence



Weak evidence ➔ Large SL ➔ sceptic has plenty of scope to "pull" findings into non-credibility

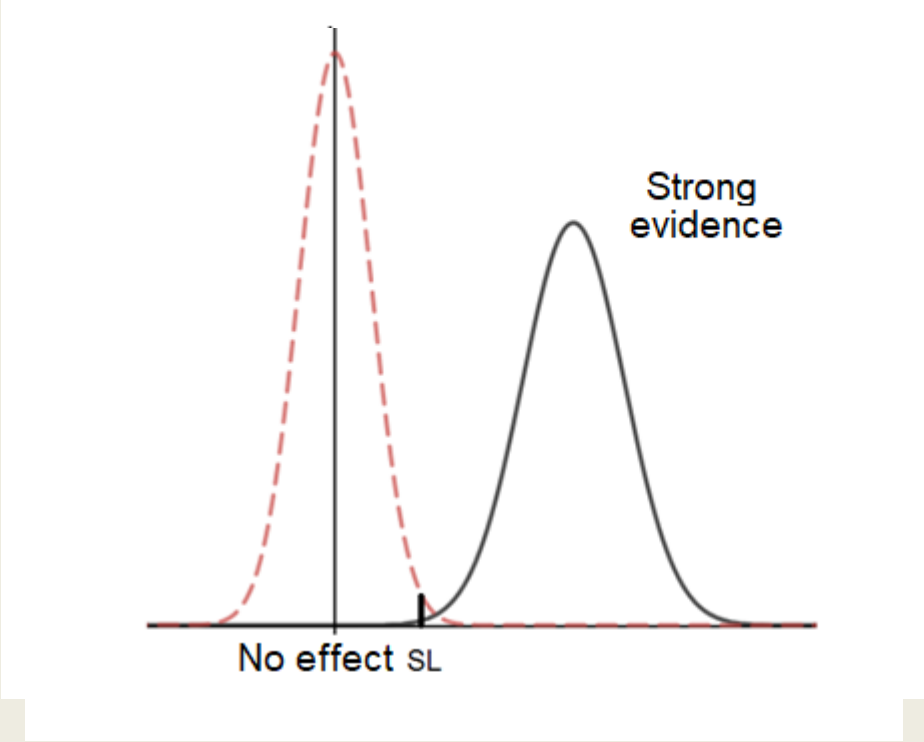
Statistically significant results

Sceptic's response to strong evidence



Statistically significant results

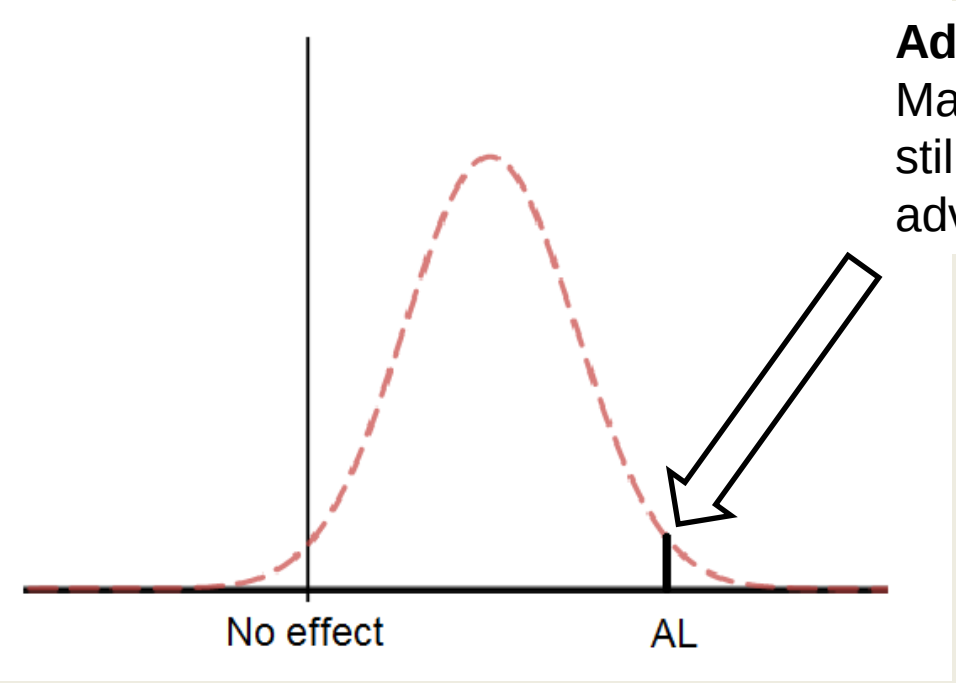
Sceptic's response to strong evidence



Strong evidence ➔ tight SL ➔ sceptic has limited ability to "pull" result into non-credibility

Statistically non-significant results

The Fair-minded Advocate

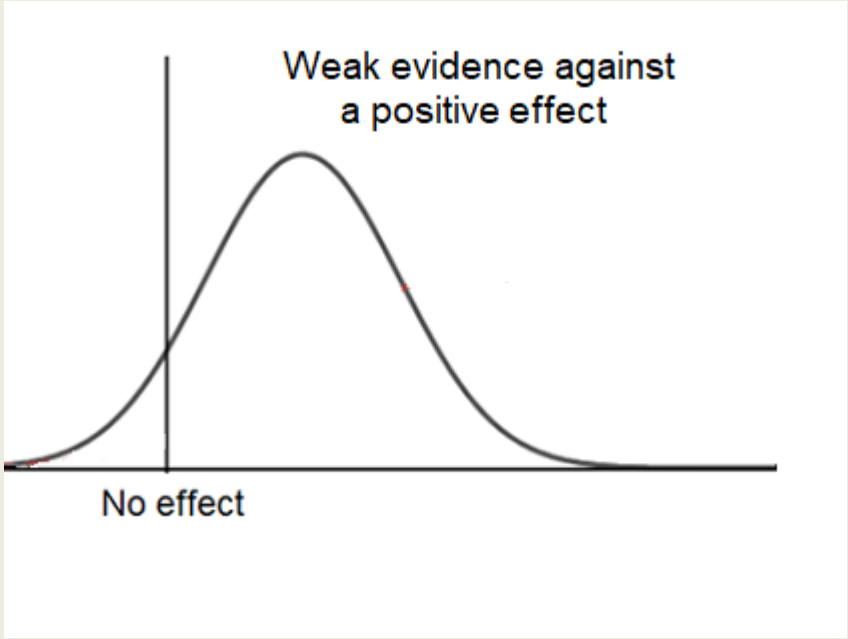


Advocacy Limit:
Maximum effect size still justifiable by advocate of effect

- 95% tails exclude no effect (“Advocacy”)
- Tails are bounded (“Fair-minded”)

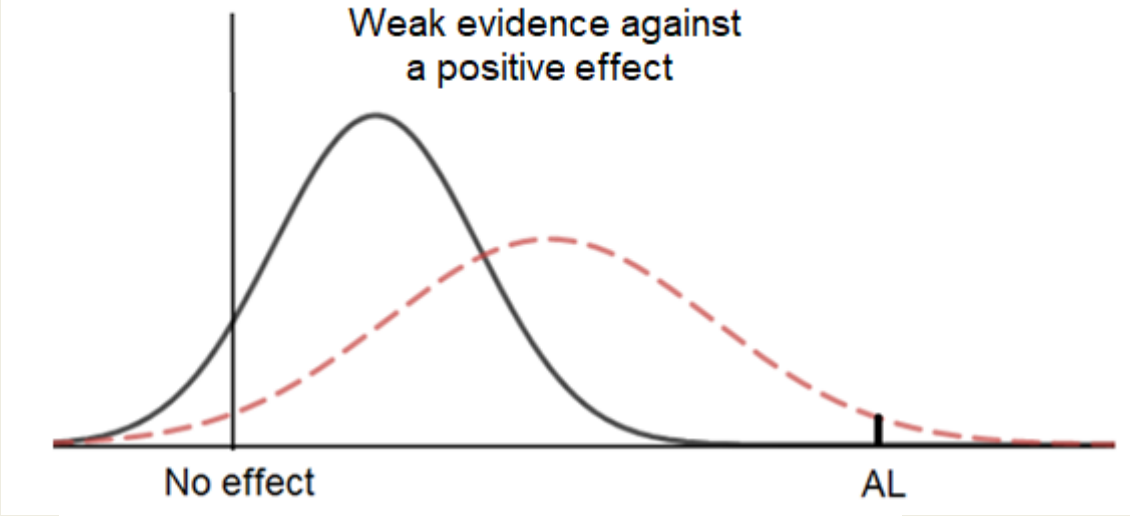
Statistically non-significant results

Advocate's response to weakly N.S. evidence



Statistically non-significant results

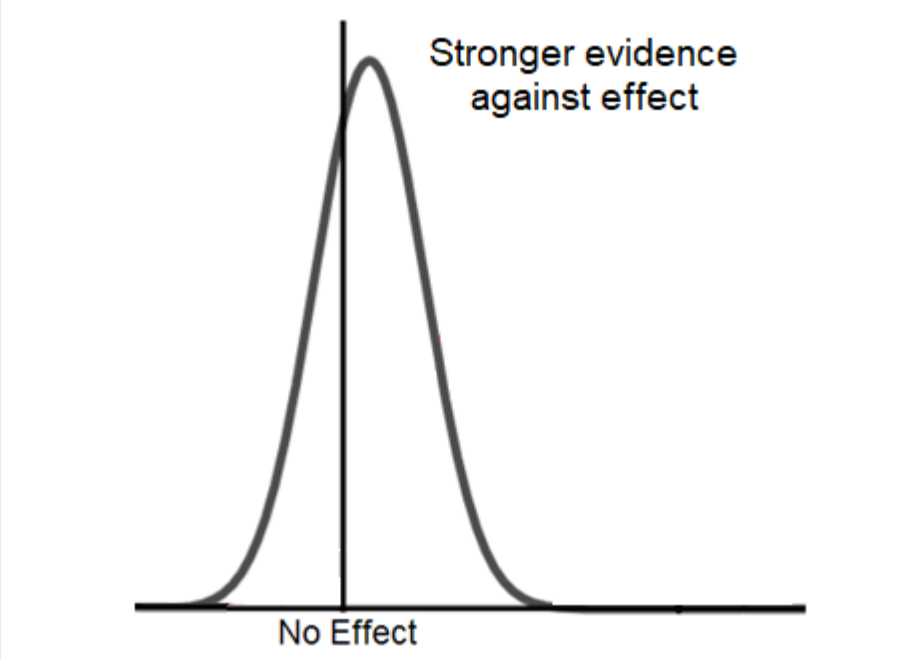
Advocate's response to weakly N.S. evidence



Weak evidence ➔ Large AL ➔ advocate has plenty of scope for "pulling" result into credibility

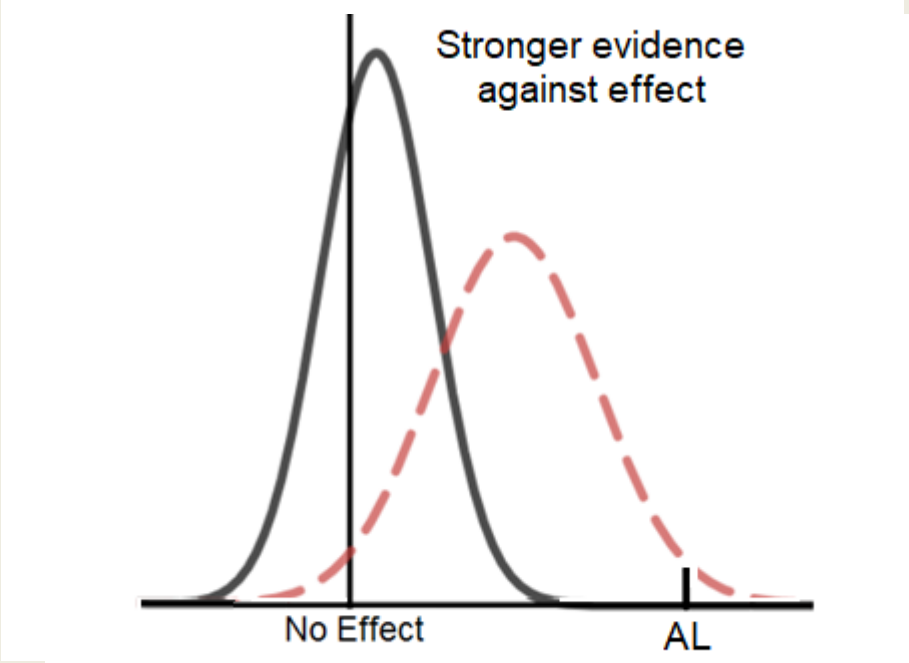
Statistically non-significant results

Advocate's response to strongly N.S. evidence



Statistically non-significant results

Advocate's response to strongly N.S. evidence



• Strong evidence against effect ➔ tighter AL ➔ advocate has much less scope for "pulling" result into credibility.

AnCred in practice

- **Input**: 95% CI summary statistic of finding
- **Analysis:** subject evidence to fair-minded challenge

Significant result:

*What level of scepticism would make result **not credible** ?*

Non-significant result:

*What level of advocacy would make result **credible**?*

- **Dichotomy:** H_1 true or false ?

AnCred in practice

- **Input**: 95% CI summary statistic of finding
- **Analysis:** subject evidence to *fair-minded challenge*:

Significant result:

*What level of scepticism would make result **not credible** ?*

Non-significant result:

*What level of advocacy would make result **credible**?*

- **Discussion:** Is level of scepticism/advocacy justifiable ?

  How does new result constrain sceptics/advocates ?

AnCred in practice

1. Getting more out of “significant” findings

“Is that really plausible ?”

Interphone study (IARC, 2010)

THEME: CANCER

Brain tumour risk in relation to mobile telephone use: results of the INTERPHONE international case–control study

The INTERPHONE Study Group*

Corresponding author. Elisabeth Cardis; CREAL, Doctor Aiguader 88, 08003 Barcelona, Spain. E-mail: ecardis@creal.cat
*List of members of this study group is available in the Appendix.

- **10,000 participants**
- **13 countries**
- **10 years,**
- **$24 million**

Interphone study

No overall glioma risk, except for heavy users:

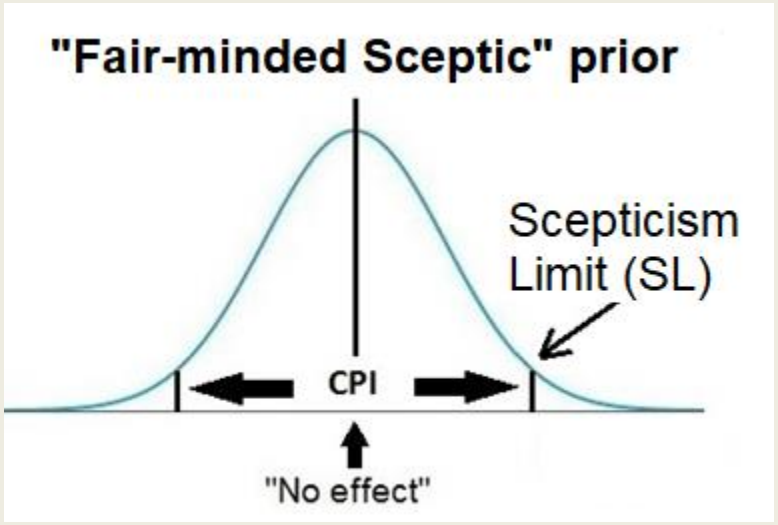
- OR 1.40; 95% CI (1.03, 1.89); p = 0.03

Challenge by fair-minded sceptic

"What prior evidence is capable of making this *not* credible at the 95% level ?"

Challenging this "significant" result

Prior: centred on no effect ("sceptical"), but tails set by strength of evidence ("fair-minded")



CPI: Critical Prior Interval for *ratios* = (*1/SL, SL*) where:

$$SL = exp\left[\frac{ln^2(U/L)}{4\sqrt{ln(U)\,ln(L)}}\right]^2$$

OR: L = 1.03; U = 1.89 ➔ SL = 2.0

Result is statistically significant, **but** is only 95% credible if prior evidence supports *at least* doubling of risk.
(It doesn't.)

AnCred in practice

3. Resolving claims of “discordant” studies

Protective effect of statins

Glioma and statins

- Good reasons/lab evidence for protective effect
- Two studies (N ~ 300-500) support it:

  - Ferris *et al* 2012: HR = 0.72 (0.52, 1.00)
  - Gaist *et al* 2013: HR = 0.76 (0.59, 0.98)

Then this happens….

“Failure to replicate”

Eur J Epidemiol (2016) 31:947–952
DOI 10.1007/s10654-016-0145-7

LETTER TO THE EDITOR

Statin use and risk of glioma: population-based case–control analysis

Corinna Seliger[1] · Christoph Rudolf Meier[2,3,4] · Claudia Becker[2] · Susan Sara Jick[3] · Ulrich Bogdahn[1] · Peter Hau[1] · Michael Fred Leitzmann[5]

*“Our findings **do not** support previous sparse evidence of possible inverse association between statin use and glioma risk”.*

N=27,000

Challenging “failure to replicate”

Two previous studies :

- Ferris *et al* 2012: HR = 0.72 (0.52, 1.00)
- Gaist *et al* 2013: HR = 0.76 (0.59, 0.98)

Seliger *et al* 2016: HR = 0.75 (0.48, 1.17)

“Failure to replicate by a very large study”

REALLY ? Wide 95% CI; similar central value…

Challenging “failure to replicate”

Two previous studies :

- Ferris *et al* 2012: HR = 0.72 (0.52, 1.00)
- Gaist *et al* 2013: HR = 0.76 (0.59, 0.98)

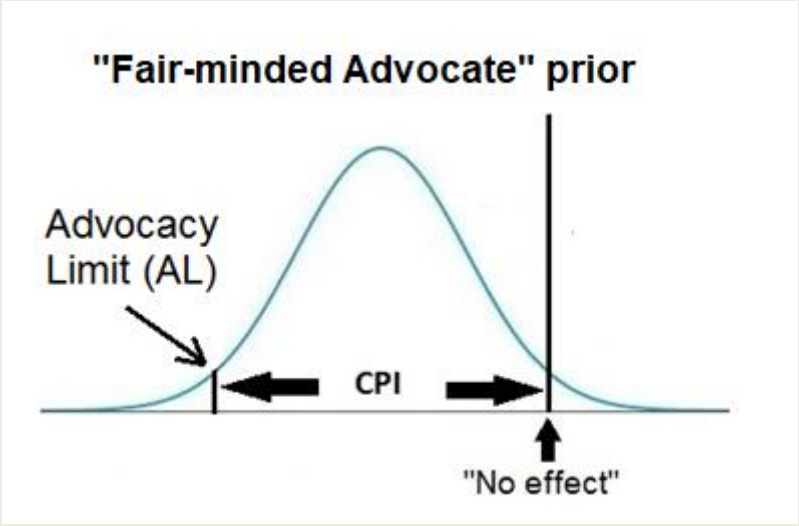
Seliger *et al* 2016: HR = 0.75 (0.48, **1.17**)

“Failure to replicate by a very large study”

REALLY ? Wide 95% CI; similar central value…

Applying AnCred

Prior: Excludes no effect (advocacy), but tails are bounded (fair-minded).



CPI: Critical Prior Interval ; for ratios = (AL, 1)

where $AL = \exp\left[-\frac{\ln(UL)\ln^2(U/L)}{2\ln(U)\ln(L)}\right]$

Seliger *et al* OR: L = 0.48, U = 1.17

➔ AL = 0.14

This **N.S.** study gives credible evidence of a **protective** effect if there is prior evidence in the range (0.14, 1.00)

➔ ENTIRELY CONSISTENT with previous studies

Despite N = 27,000, Evidence is *weak* (broad CI *and* CPI)

AnCred in practice

4. Avoiding over-interpretation of studies

Resuscitation in septic shock

Is CRT better marker than serum lactate ?

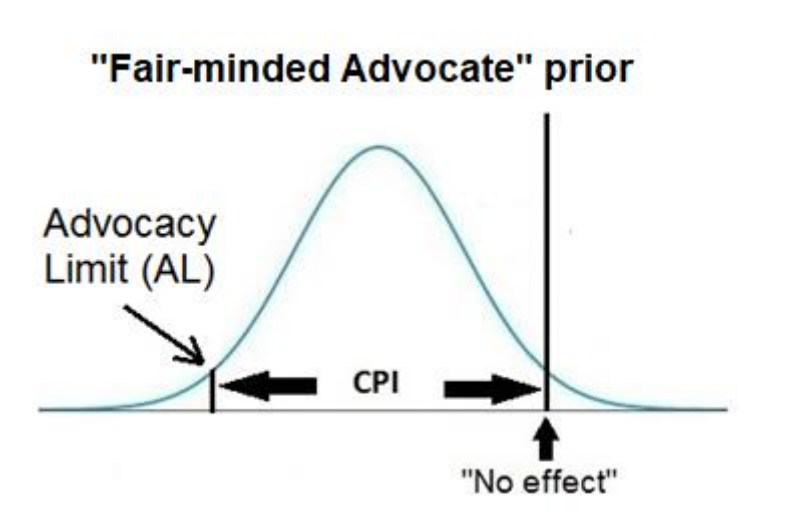
Hernandez et al *JAMA* 2019: ANDROMEDA-SHOCK RCT (N = 424)

Mortality CRT v SL: HR = 0.75 (0.55, 1.02)
Mortality risk difference: -8.5% (-18.2, +1.2)

➔ CRT "…did **not** reduce all-cause mortality"

Challenging the "non-significance"

Prior: Excludes no effect (advocacy), but tails are bounded (fair-minded).



CPI: Critical Prior Interval ; for ratios is (AL, 1) where

$$AL = \exp\left[-\frac{\ln(UL)\ln^2(U/L)}{2\ln(U)\ln(L)}\right]$$

Hernandez *et al* HR: L = 0.55, U = 1.02 ➔ AL < 0.01

This N.S. study provides credible evidence of a **protective** effect if there is ANY prior support for one

➔ Encouraging, and bigger studies certainly merited

Challenging editors/reviewers

Authors did NOT want to focus solely on non-significance

"[W]e think CRT is better than lactate"

BUT

"Reviewers & editor asked us to temper our enthusiasm and stick to the cold stats"

AnCred gives researchers quantitative alternative to "pass/fail" dichotomy

AnCred in practice

Conclusions

AnCred: one small step, but easily taken

- Familiar input (CIs); readily interpreted output
- Extracts more from summary statistics
- Helps promote publication of “null” results
- Highlights weak evidence from large studies
- Replaces “dichotomania” with contextual debate

AnCred developments

- Analysis of “out of the blue” findings via *intrinsic credibility* (Matthews 2018, Held 2019)
- Replication probability (Held 2019, 2020)
- Beyond the Normal distribution, inferences on differences and ratios

Easily applied retrospectively

Inferential issues addressed by AnCred

Replication “failures”

“Absence of evidence = evidence of absence”

Implausible claims

Underpowered studies

Borderline significance/non-significance

“Out of the blue” studies

Happy hunting !

Thank you

rajm@physics.org

References

- Matthews RAJ 2019 Moving Towards the Post $p < 0.05$ era via the Analysis of Credibility *Am Stat* **73** *202-212*

- Held, L 2019 The assessment of intrinsic credibility and a new argument for $p < 0.005$ *Roy Soc Open Sci*

- Matthews RAJ 2017 Beyond "significance": principles and practice of the Analysis of Credibility *Roy Soc Open Sci*

- Matthews RAJ 2001. Why should clinicians care about Bayesian methods? J Stat Plan Inf 2001 Mar 1;94(1):43-58.

- Spiegelhalter DJ, Abrams KR, Myles JP. 2004 *Bayesian approaches to clinical trials and health-care evaluation*. (Chichester: Wiley & Sons) 75 *et seq*

Appendix: Summary of basic AnCred formulas

For 95% CIs (L, U) for diffs in means/proportions ~ N[μ, ϕ]

Significant results: if prior evidence exists for effects *outside* Critical Prior Interval (CPI) of *(-SL, +SL)* where for differences of means/proportions expressed as CIs of (L, U):

$$SL = (U - L)^2/4vUL$$

➔ Evidence for a real effect is also **credible** at 95% level.

Non-significant results: if prior evidence exists for (positive) effects *inside* CPI of *(0, AL)* where for differences of means/proportions expressed as CIs of (L, U):

$$AL = -(U + L)(U - L)^2/2UL$$

➔ There is still credible evidence for a real effect at 95% level.

For ratios (OR, HR – *not* RRs), SL and AL follow from $SL \Rightarrow Ln(SL)$ *etc.*
For full derivations see Matthews (2017)