



A Practical Guide to Interpreting the Research Literature

How to Detect Flaws, Avoid Deception

Austin Parish, MD MS

Disclosures Disclaimers & Ethos

Disclosures: I have no disclosures or conflicts of interest.

Disclaimer: I cannot teach you everything about interpreting research in 20 minutes.

Will not cover the legal side of practice – follow guidelines and local standard of care
Will not cover raw statistics – not defining t-tests, etc.



Academic Emergency Medicine
A GLOBAL JOURNAL OF EMERGENCY CARE

ORIGINAL CONTRIBUTION

An umbrella review of effect size, bias, and power across meta-analyses in emergency medicine

Austin J. Parish MD, MS ✉, Denley M. K. Yuan MD, Jason R. Raggi DO, Oluyemi O. Omotoso MBBS, Jason R. West MD, John P. A. Ioannidis MD, DSc



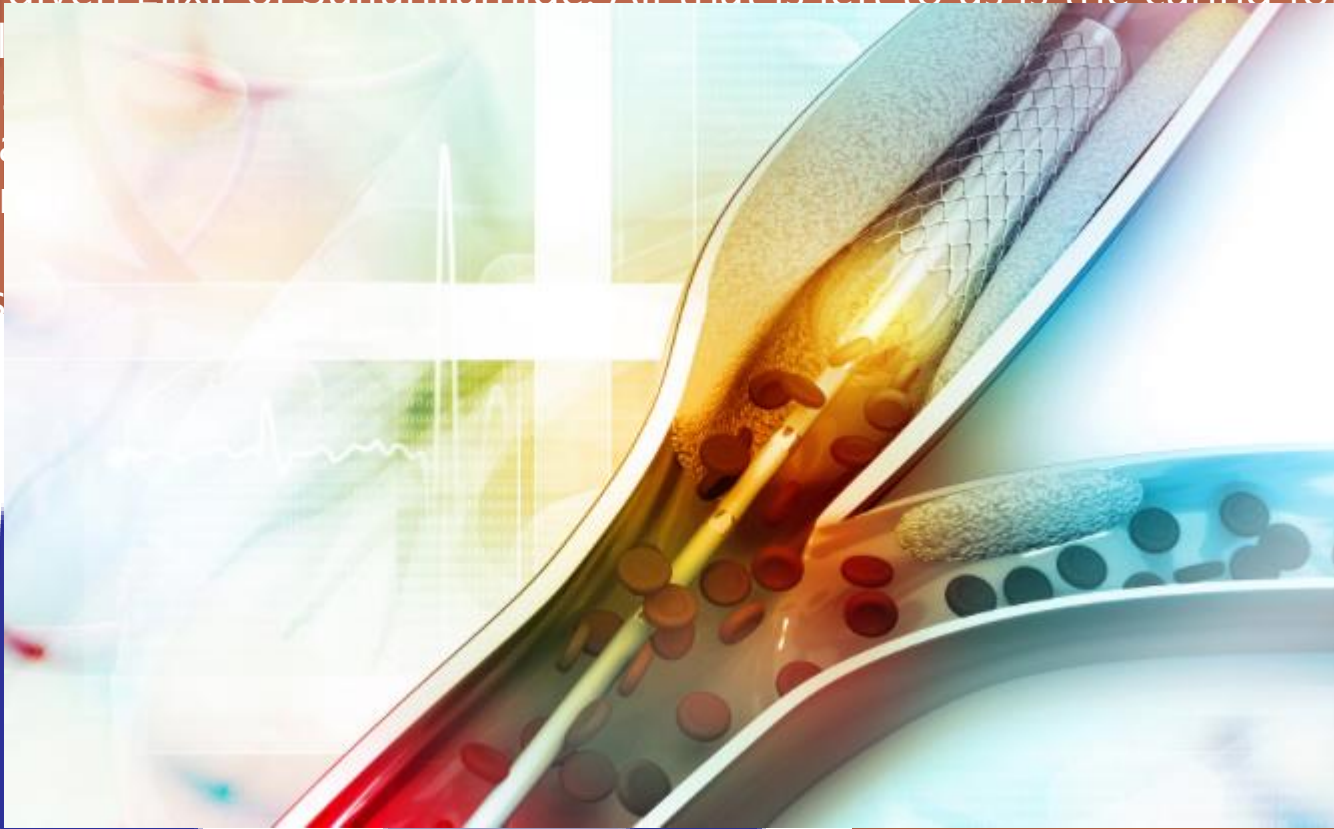
Medicine Based on *Evidence*

1937:

"The first time I ever had occasion to call in a doctor for her... she was given Flixir of Sulfanilamide. All that is left to us is the caring for her."

her
can
scre
insan

"It is
wi
ou



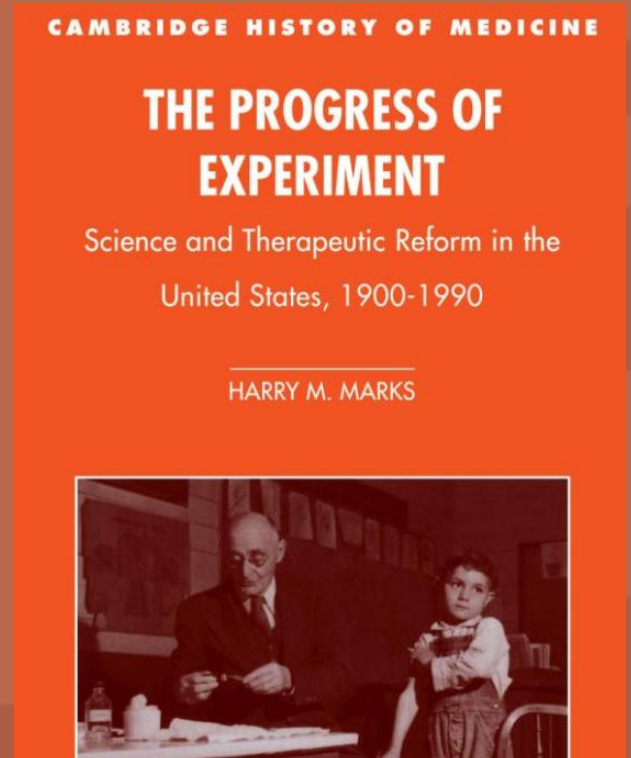
Medicine Based on *Evidence*

The first clinical trial was carried out in 1747
James Lind: n = 12 scurvy patients

"On the 20th of May 1747, I selected twelve patients in the scurvy... Their cases were as similar as I could have them. They all in general had putrid gums, the spots and lassitude, with weakness of the knees. They lay together in one place, being a proper apartment for the sick in the fore-hold; and had one diet common to all...

Two were ordered each a quart of cyder a day. Two others took twenty-five drops of elixir vitriol three times a day ... Two others took two spoonfuls of vinegar three times a day ... Two of the worst patients were put on a course of sea-water ... Two others had each two oranges and one lemon given them every day ... The two remaining patients, took ... an electary recommended by a hospital surgeon ...

The consequence was, that the most sudden and visible good effects were perceived from the use of oranges and lemons; ...of those who had taken them, being at the end of six days fit for duty."



Medicine Based on *Rationality*

Physicians aren't great at understanding research

42% correctly answer a question about p-values

26% correctly answer a question about interpreting diagnostic test

12% got both correct

Science is an extension of rationality

· requires reasoning in an uncertain world

Probability is most usefully thought of a measure of **strength of belief**, that we **update in response to evidence**

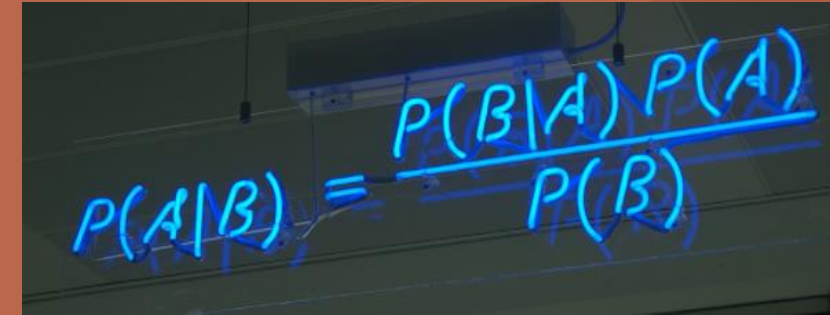
Bayes theorem (update a flawed model of the world based on incomplete evidence)

Evidence can be very strong, and thus update our prior belief a lot

Evidence can be weak and not move our belief at all

Cromwell's rule

("I beseech you, in the bowels of Christ, think it possible that you may be mistaken")


$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



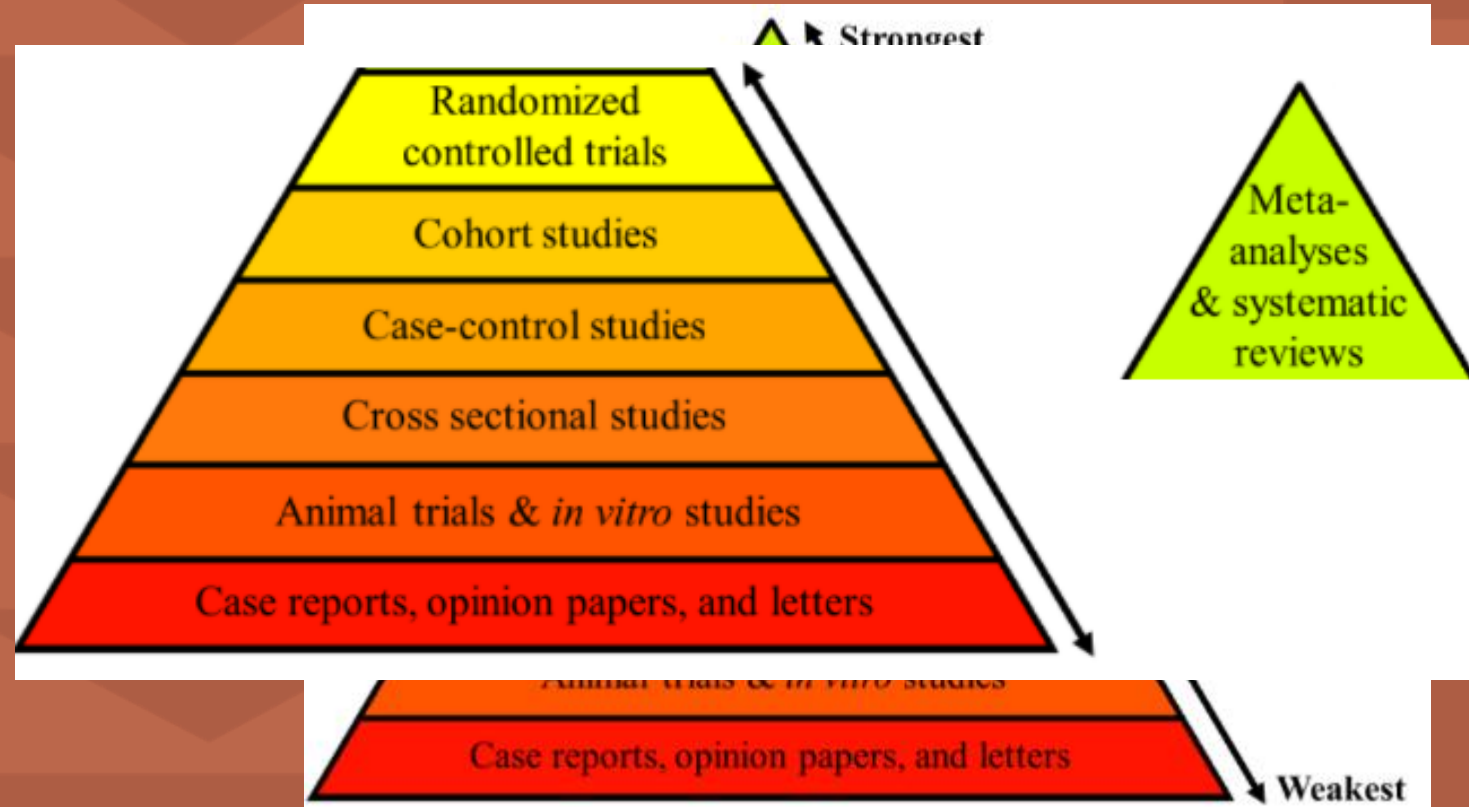
Randomization, Causality and the RCT

“Levels of evidence”

RCTs are *different*; the only* way to truly determine causality is randomization

(Essentially) no other process can eradicate confounding

Meta-analyses combining flawed, biased, broken studies are meaningless



Primary and Secondary Outcomes

#1: What outcome was this trial designed to detect?

Primary outcome: the outcome that the trial was *designed* and *powered* to study

Secondary outcome: less important outcomes but still considered worthy of study;
do *not* drive the design/sample size of the trial

Example:

Primary outcome: **hospitalization**

Secondary outcomes: all-cause mortality, time to death, time to clinical improvement, number of days w/ symptoms...



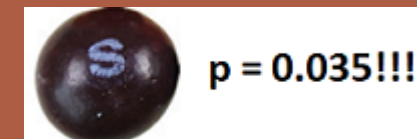
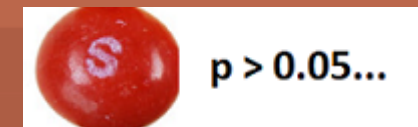
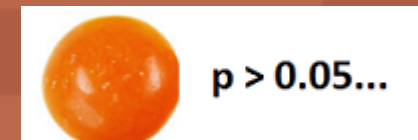
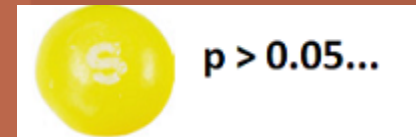
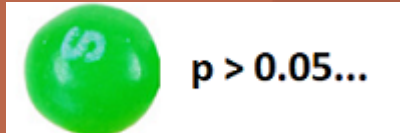
THE LANCET
Global Health

ARTICLES | ONLINE FIRST

Effect of early treatment with fluvoxamine on risk of emergency care and hospitalisation among patients with COVID-19: the TOGETHER randomised, platform clinical trial

Gilmar Reis, PhD • Eduardo Augusto dos Santos Moreira-Silva, PhD • Daniela Carla Medeiros Silva, PhD • Prof Lehana Thabane, PhD • Aline Cruz Milagres, RN • Thiago Santiago Ferreira, MD • et al. [Show all authors](#) • [Show footnotes](#)

[Open Access](#) • Published: October 27, 2021 • DOI: [https://doi.org/10.1016/S2214-109X\(21\)00448-4](https://doi.org/10.1016/S2214-109X(21)00448-4) •



Patient-Centered, Subjective and Surrogate *Outcomes*

#1: What outcome was this trial designed to detect?

Patient-centered: outcomes that matter to patients

(Patients may care more about days lived well than total number of days)
(Quality of life, mortality, disability free days...)

Subjective: outcomes that depend on subjective judgment (of patient or provider)

Patient-centered outcomes can be subjective (quality of life) or objective (mortality)
More subject to distortion if not clear blinding, "novelty bias"

Surrogate: outcomes that are (hopefully) associated with the outcomes we really care about, but are easier to measure (occur more frequently, less rare)

- Often drugs adopted based on surrogate outcomes ultimately proven not to work, or worse (clofibrate, doxazosin, flecainide...)



Surrogate marker examples

- Cholesterol levels for heart disease
- Bone density for fractures
- Hemoglobin A1c for diabetes
- Progression-free survival for cancer

Placebo vs Active Controls

#2: What is the *control group*?

Randomize patients to an experimental (intervention) group, and a **control group**

Control group can be either **placebo** or **active control**

Placebos: **strong medicine**

~25% of patients taking a placebo report significant adverse effects

Patients that *know* they are receiving a placebo are significantly helped (IBS: 60% improvement in symptoms; low-back pain 28% reduction in pain; cancer fatigue 29% improvement in fatigue... (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6889847/>))

For new drugs/interventions in areas with *established treatments*, we should be comparing to an **active control**



Blinding and Allocation Concealment

#3: Is the trial sufficiently **blinded**? Is the sequence of allocation appropriately **concealed**?

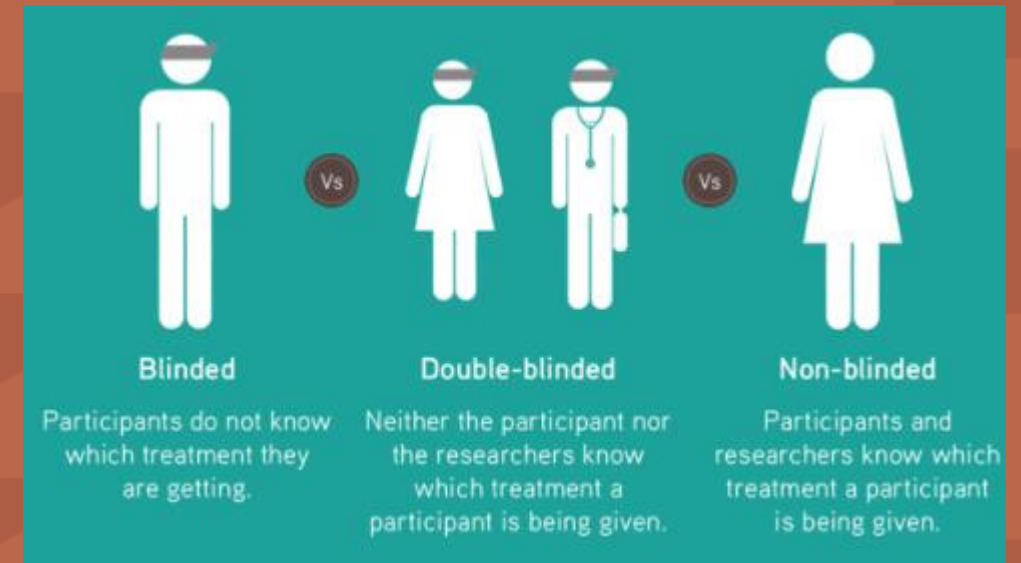
Does it matter?

Lack of clear double-blinding exaggerated results by **22%**

Lack of clear allocation exaggerated results by **15%**

This result was for outcomes w/ subjective components

Only 3 to 7% for objective outcomes (mortality, etc.)



Annals of Internal Medicine®

Search Journal

LATEST ISSUES IN THE CLINIC JOURNAL CLUB MULTIMEDIA CME / MOC AUTHORS / SUBMIT

Research and Reporting Methods | 18 September 2012

Influence of Reported Study Design Characteristics on Intervention Effect Estimates From Randomized, Controlled Trials

Jelena Savović, PhD, Hayley E. Jones, PhD, Douglas G. Altman, DSc, Ross J. Harris, MSc, ... [View all authors +](#)

[Author, Article and Disclosure Information](#)

<https://doi.org/10.7326/0003-4819-157-6-201209180-00537>

Against p-values

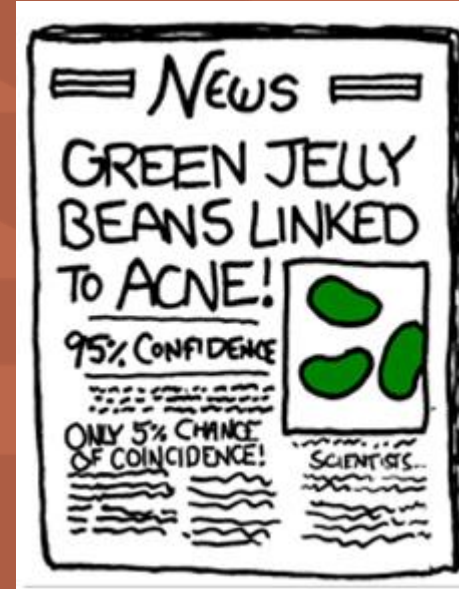
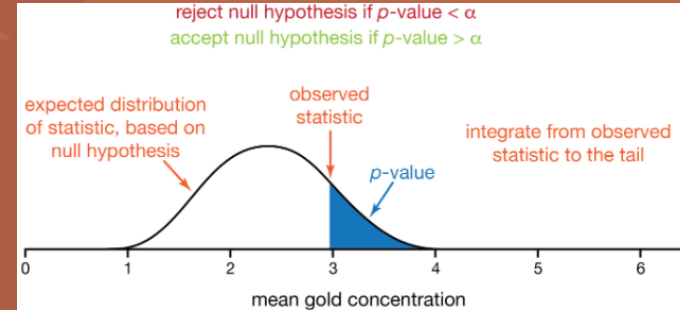
"This is the calculated number that shows what the chance actually is that the data supports your hypothesis..." (Mel Herbert et al, *EMRAP*)

"The p-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct" (Wikipedia)

Often, p-values are the "threshold" for "good (publishable) research"

P-hacking

Clinical vs statistical significance



1. Stop collecting data once $p < .05$
2. Analyze many measures, but report only those with $p < .05$.
3. Collect and analyze many conditions, but only report those with $p < .05$.
4. Use covariates to get $p < .05$.
5. Exclude participants to get $p < .05$.
6. Transform the data to get $p < .05$.

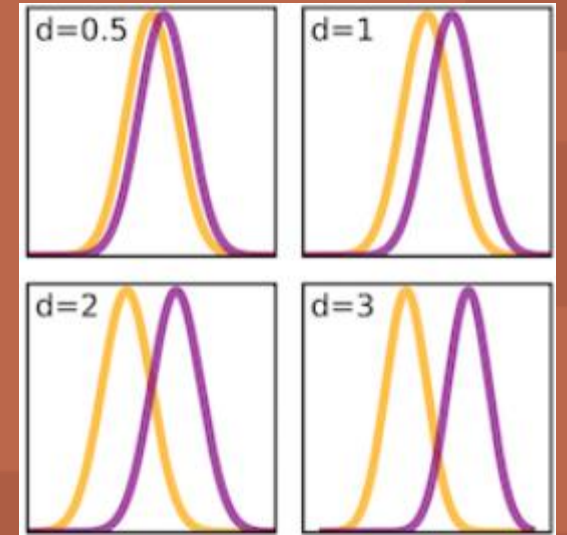
Effect Sizes and Confidence Intervals

Effect sizes are more important than p-values

Effect sizes are **measures** of the magnitude of the difference between two groups, for an **outcome** of importance

Effect sizes should be reported with a **confidence interval**, or even better, a **credible interval**

Some interventions are so large that they do not need randomized evidence (!!)



Clinical vs Statistical Significance

#4: If the outcome is positive, does it *matter*?

Statistical significance: monetized, prioritized, addictive, deceptive, *common**

Clinical significance: meaningful, rare

BEWARE FALSE CONCLUSIONS

Studies currently dubbed 'statistically significant' and 'statistically non-significant' need not be contradictory, and such designations might cause genuine effects to be dismissed.

Table 1. Key Questions to Ask When the Primary Outcome Is Positive.

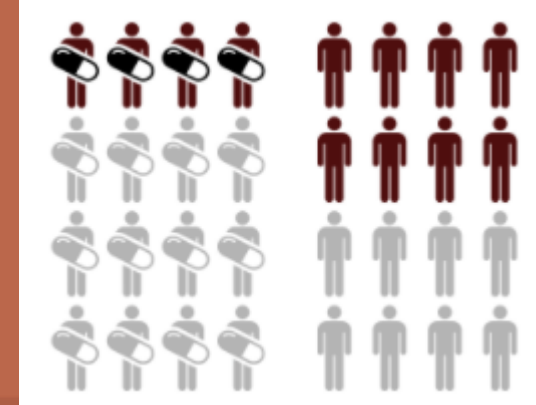
- Does a P value of <0.05 provide strong enough evidence?
- What is the magnitude of the treatment benefit?
- Is the primary outcome clinically important (and internally consistent)?
- Are secondary outcomes supportive?
- Are the principal findings consistent across important subgroups?
- Is the trial large enough to be convincing?
- Was the trial stopped early?
- Do concerns about safety counterbalance positive efficacy?
- Is the efficacy–safety balance patient-specific?
- Are there flaws in trial design and conduct?
- Do the findings apply to my patients?



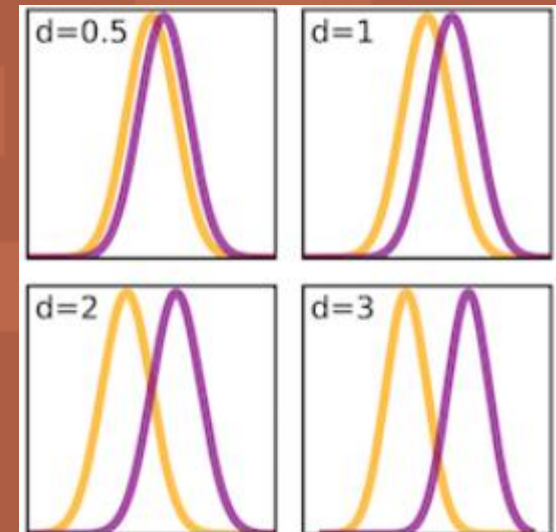
Deception by Effect Choice (1)

The difference in outcome between groups can be summarized in different ways:

Risk ratio
Odds ratio
Risk difference
Number needed to treat/harm...



- 1) Plain language summaries of ratio results often lead to **confusion**
- 2) Odds ratios and risk ratios are **very different**, especially if the outcome occurs frequently
- 3) Risk differences should **always** be reported
- 4) Number needed to treat/harm gives an **intuitive** understanding of the benefit (or risks) of an intervention



Deception by Effect Choice (2)

EXAMPLE: *Intervention:* **Corticosteroids in sore throat**

Outcome: Complete resolution of pain at 24 hours

Placebo: 158/1000 (15.8%)

Steroids: 379/1000 (37.9%)

Risk ratio = $37.9/15.8 \sim 2.40$

Odds ratio = $(37.9/62.1) / (15.8/84.2) \sim 3.30$

Absolute % increase: $37.9 - 15.8 \sim 22\%$

NNT = $(100\% / 22.1\%) \sim 5$



Deception by Effect Choice (3)

sampling weight. The risk of hypertension was synergistically and significantly increased among persons with both insomnia or poor sleep, and short sleep duration. The presence of both insomnia and an objective sleep duration ≤ 5 h increased the risk for hypertension by about 500% (OR = 5.12., 95% CI 2.2–11.8) compared to the group without insomnia/poor sleep complaint

EXAMPLE:

Actual data:

Normal sleeping + duration >5 hours: **25% have HTN**

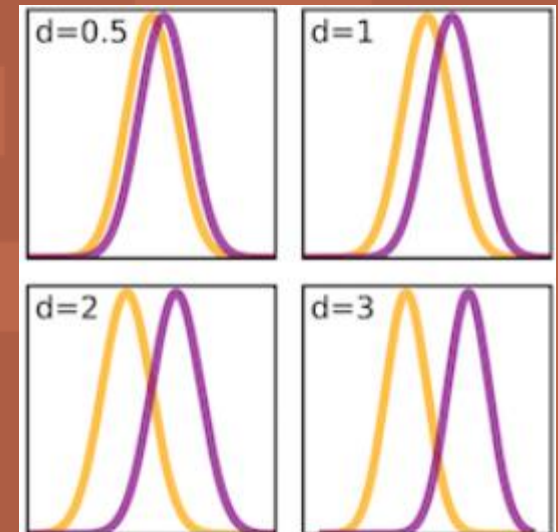
Insomnia + duration <5 hours: **63% have HTN**

Odds ratio = $(63\% / 37\%) / (25\% / 75\%) \sim 5.1$

Risk ratio = $63\% / 25\% \sim 2.5$

Absolute risk difference = $63\% - 25\% = 38\%$

The actual "risk increase" was only 38%!!!



Sample Size and Power

If the **true** effect you are looking for is very large, you don't need many patients!

Unfortunately, we live in a field of small effects (median OR = 0.70, Cohen's $d = -0.2$: small)

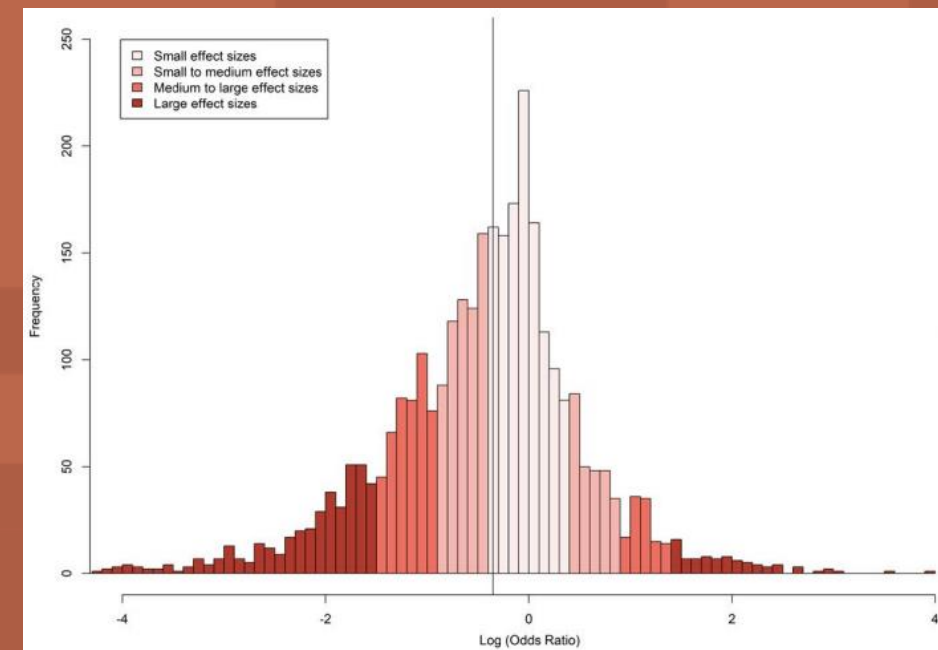
This relationship is **critical**, especially for novel effects

Early and small demonstrations of treatments are more likely to cross the "p-value threshold" if they **randomly** find a very large effect

Early effects are often substantially larger than the "true" effect



Surprising, new, and small – be wary!



Judging Clinical Trials: Five Questions to Ask

#1: How surprising is this result to me, based on my previous knowledge?

#2: Are the outcomes reported meaningful to patients?

#3: Is there a clinically significant *effect size*, in a meaningful measure such as NNT?

#4: Is the trial large enough to detect the effect it reports?

#5: Is the control group appropriate: for a novel therapy in a field with accepted treatments, the comparison should be active!

#6: Especially if the outcome is subjective, was there adequate **double blinding** and **allocation concealment**?

#7: Is this an early, unparalleled result?

Early effects are often substantially larger than the "true" effect

Judging Systematic Reviews and Meta-Analyses (SRMAs)

SRMAs: increased **2700%** from 1991 to 2015

Each year, more SRMAs published than new trials

Huge amount of redundancy (67% of meta-analyses published in 2010 had another published on same topic within one year)

From 2008 to 2015:

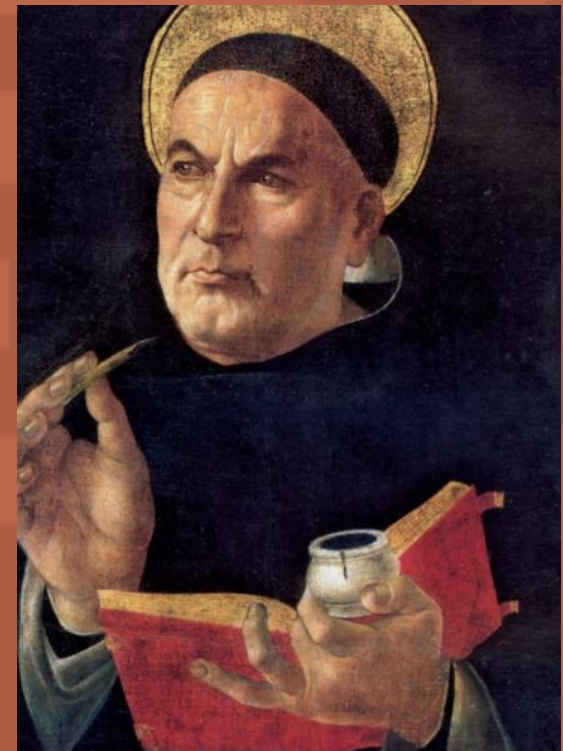
21 meta-analyses of statins to prevent Afib after cardiac surgery

Meta-analyses on same topic can reach different conclusions

Framing, choice of studies (done retrospectively)

"Systematic" is *not* systematic

~70% of SRMAs in Critical Care Literature had one or more major flaws



Judging Systematic Reviews and Meta-Analyses (SRMAs)

Questions to ask about SRMAs:

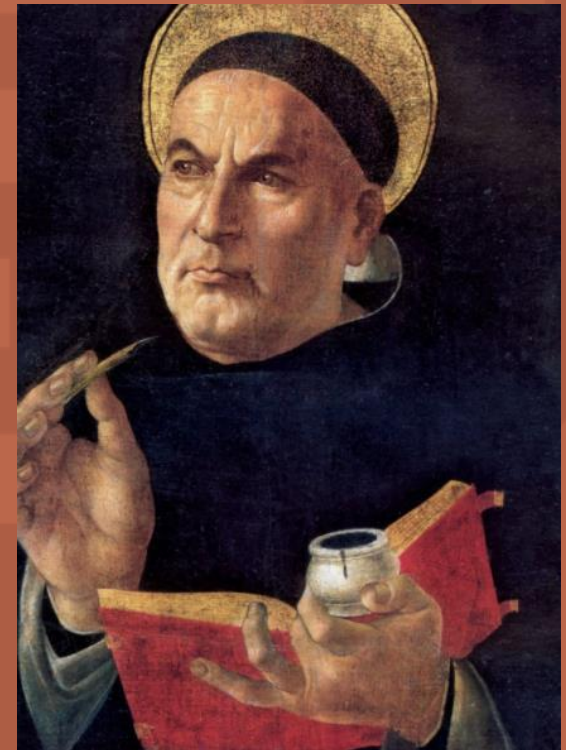
Was the search strategy comprehensive?

Searched at least 2 databases'

Searched trial registries

Searched grey literature

Hand searched reference lists of included studies



EBM: Possible, Difficult, Vital

Are guidelines/physician consensus/society statements the top level of EBM?
Only 33% of guidelines used **systematically synthesized evidence**

"We don't reach agreement when we have discovered the truth. Instead, we have discovered the truth when we reach agreement." (Giani Vattimo)

"The world does not speak. Only we do." (Richard Rorty)



Questions?