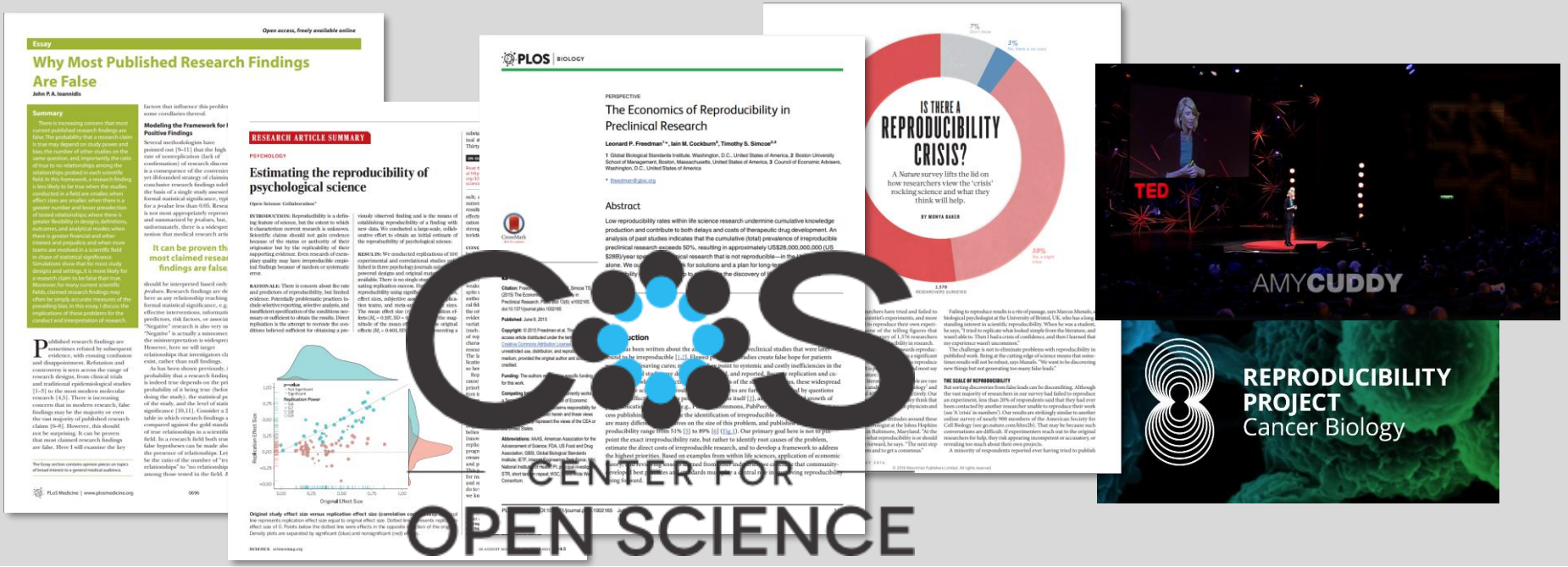
Replication crisis in science

anything to do with clinical research?

Sven Trelle, CTU Bern

April 27, 2022



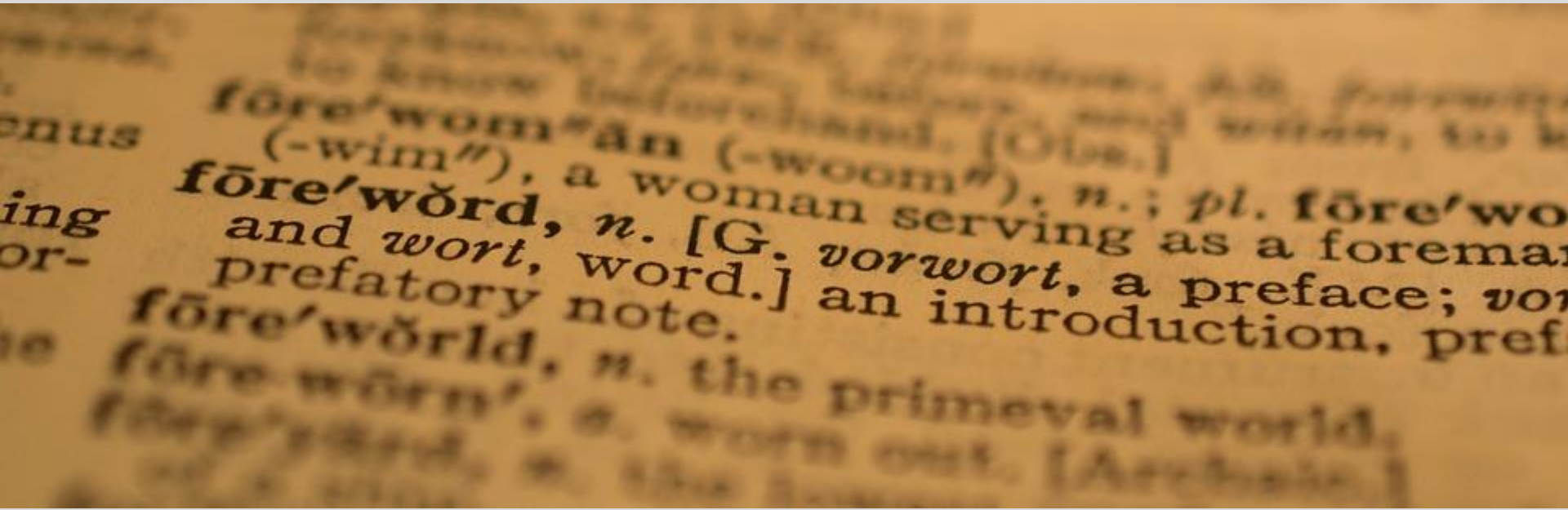
What I will cover

The menu

- Some definitions
- Replication crisis
- Voting on confidence/(un)certainty
- Clinical research

Some definitions

No details …



Terminology

UNIVERSITÄT BERN

- Reproducibility (Quality Control)
- Replicability (Quality Assurance)
- Robustness (sensitivity) (Quality Assurance)
- Generalizability

		Data	
		Same	Different
Analysis (code)	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Kirstie Whitaker (https://youtu.be/NDNYPDm1-2c)

Reproducibility

Terminology from another perspective

UNIVERSITÄT BERN

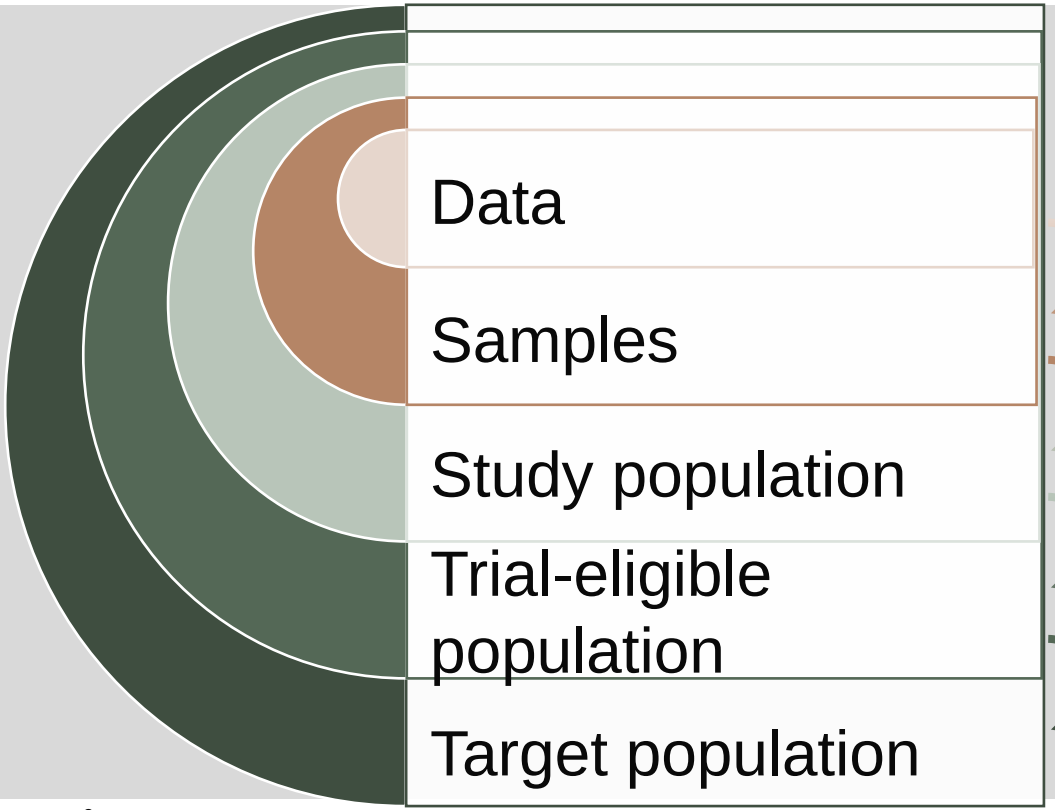
- Methods reproducibility
  - Study can be/is exactly* (?) repeated
- Results reproducibility
  - Same (?) results from an independent (closely matched) study
- Inferential reproducibility
  - Drawing qualitatively the same conclusions from an independent analysis or study

* Too exactly would actually be useless …

Goodman 2016

Induction

The basic idea of empirical (clinical) research



Measurement
- Data quality
- Reliability
- Construct validity

Internal validity
- Randomization

Generalizability
- Reflecting the real world

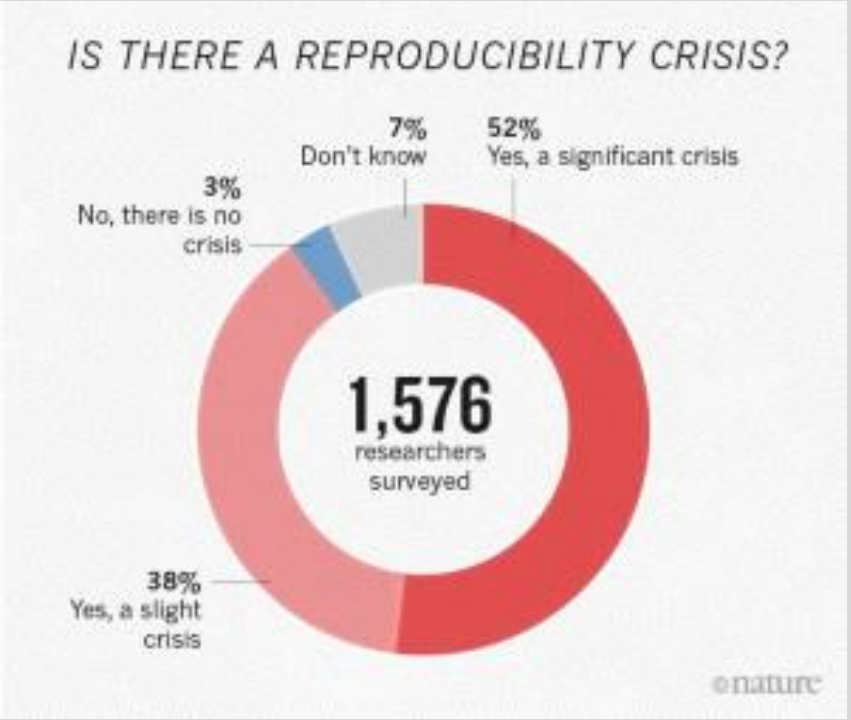
Transportability
- Reflecting the real world and extending to the real world

Adapted from Spiegelhalter D 2019 (see also Dahabreh IJ et al. 2020)

Replication crisis

in science

The replication crisis (also called the replicability crisis and the reproducibility crisis) is an ongoing methodological crisis in which it has been found that the results of many scientific studies are difficult or impossible to reproduce. Because the reproducibility of empirical results is an essential part of the scientific method,[2] such failures undermine the credibility of theories building on them and potentially call into question substantial parts of scientific knowledge. *Wikipedia (25.04.2022)*



Baker 2016

How many studies are not replicable?

The simple, empirical question

UNIVERSITÄT BERN

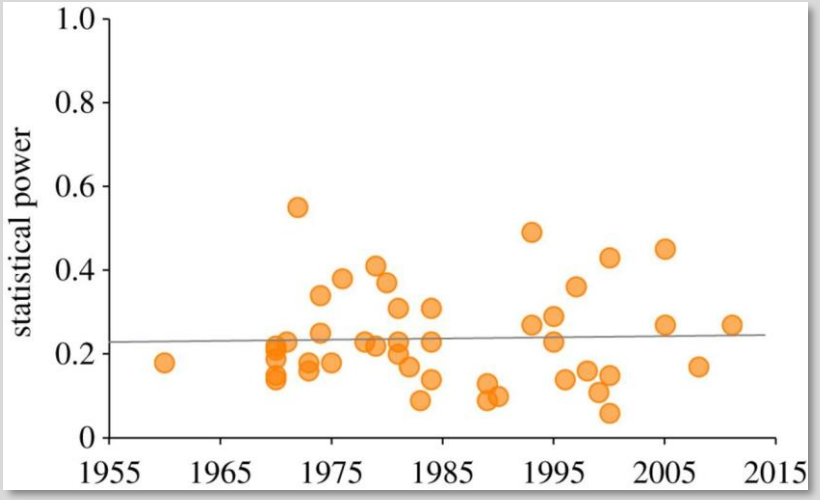
- Psychology (Open Science Collaboration 2015)
  - *The mean effect size (r) of the replication effects (Mr = 0.197, SD = 0.257) was half the magnitude of the mean effect size of the original effects (Mr = 0.403, SD = 0.188), representing a substantial decline. Ninety-seven percent of original studies 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;*
- Social sciences (Camerer 2018)
  - We find a significant effect in the same direction as the original study for 13 (62%) studies, and the effect size of the replications is on average about 50% of the original effect size. Replicability varies between 12 (57%) and 14 (67%) studies for complementary replicability indicators.
- Preclinical research (Freedman 2015)
  - *An analysis of past studies indicates that the cumulative (total) prevalence of irreproducible preclinical research exceeds 50%,*
- Clinical medicine (Ioannidis 2005)
  - *Of 49 highly cited original clinical research studies, 45 claimed that the intervention was effective. Of these, 7 (16%) were contradicted by subsequent studies, 7 others (16%) had found effects that were stronger than those of subsequent studies, 20 (44%) were replicated, and 11 (24%) remained largely unchallenged.*

One explanation: low statistical power

Social, behavioural, biological sciences

UNIVERSITÄT BERN

- 19 reviews (1992 to 2014)
- Power to detect small effects (*d*=0.2): *the kind most commonly found in social science research*

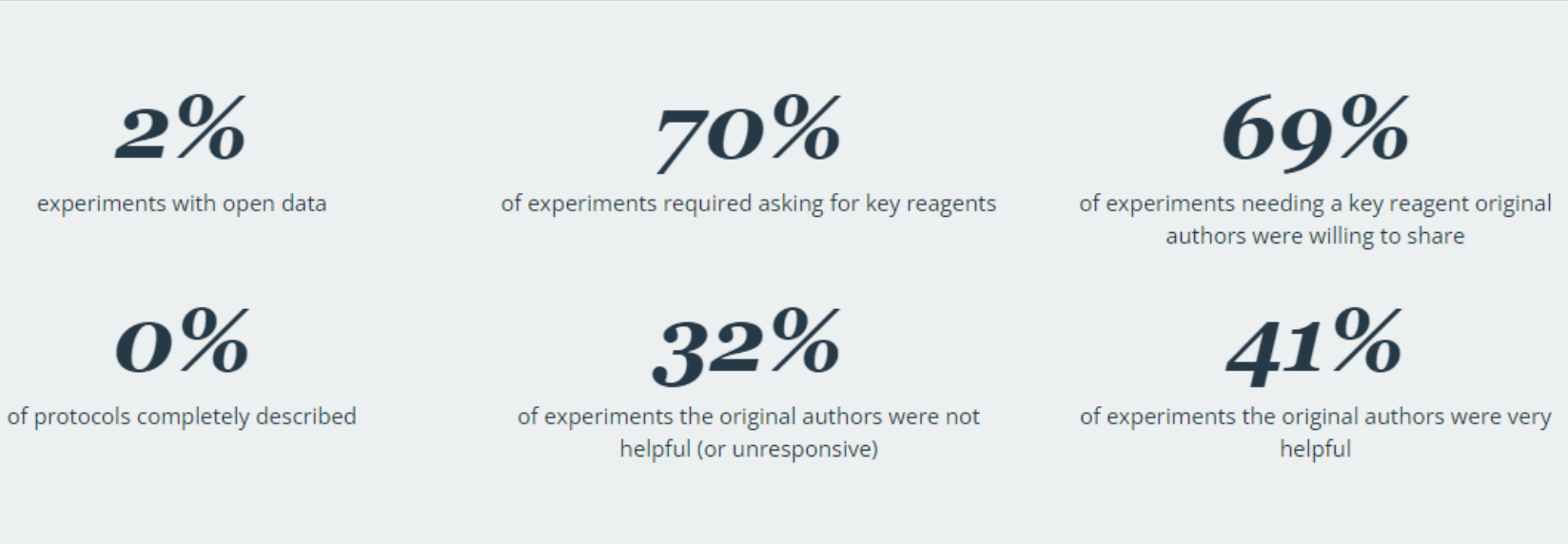


Smaldino 2016

Reproducibility Project

Cancer Biology

– 193 experiments from 53 papers



https://www.cos.io/rpcb?hsLang=en; Errington 2021

Reproducibility Project

Cancer Biology

– 50 replications from 23 papers (158 effects)

– Replication effect sizes were 85% smaller on average

– Original positive results were half as likely to replicate successfully (40%) than original null results (80%)

https://www.cos.io/rpcb?hsLang=en

Have replication rates changed over time?

Decreased or increased

– According to David Jensen

To my knowledge, we don't have good evidence on this question

The interpretation of the answer would also depend on whether we believe that research questions have become easier or more difficult and whether the underlying technologies for research have improved

"Crisis" implies urgency and recency, but we don't appear to have evidence for this

David Jensen (https://www.umassmed.edu/globalassets/ccts/jensen-reproducibility-talk.pdf)

Voting

Your spontaneous thoughts

- You will be presented with results from true (some outdated!) clinical trials and vote how confident/certain you are about the observed effect.
- General assumptions:
  All trials are
  - Ethical
  - Methodologically sound
  - Well powered
  - Measure patient-relevant outcomes

Question #1

Chondroitin

UNIVERSITÄT BERN

A randomized-controlled trial compared chondroitin (medication) with placebo to treat osteoarthritis pain. The trial shows that chondroitin reduces pain on average by 2.14 standard deviation units (Cohen's *d*; assume that 0.3 to 0.5 is a clinically relevant effect). The 95% confidence interval ranges from 1.49 to 2.80 (0 means no benefit of chondroitin). You know that a joint replacement reduces pain on average by 1.0 standard deviation units.

How confident/certain are you about this observed effect:

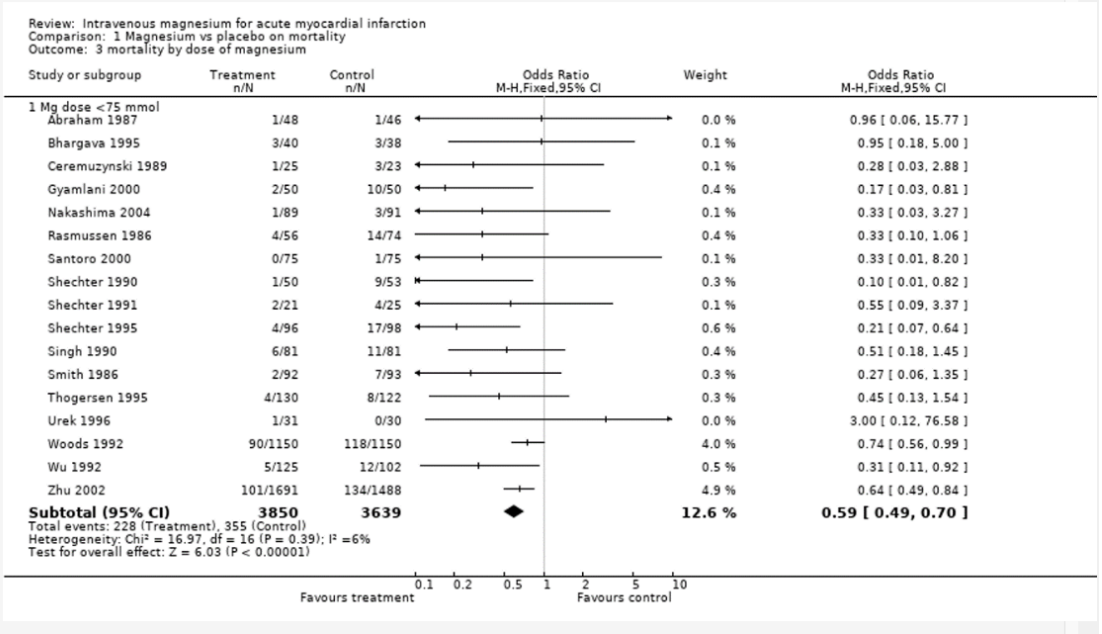
- **A** I am sure that chondroitin reduces pain
- **B** I am somehow/pretty convinced that chondroitin reduces pain
- **C** I do not know whether chondroitin reduces pain
- **D** I am somehow/pretty convinced that chondroitin does not reduce pain
- **E** I am sure that chondroitin does not reduce pain

Question #2

Magnesium

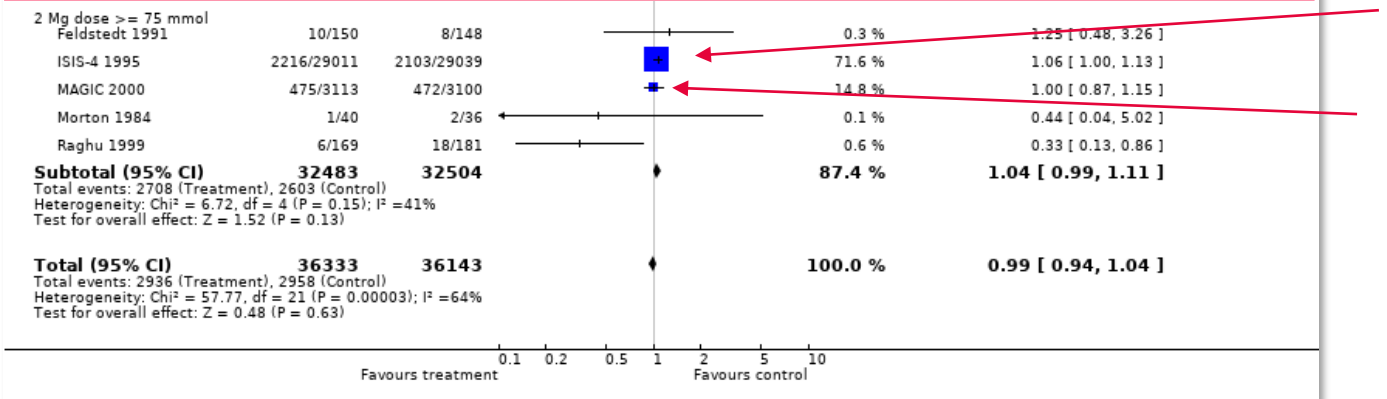
UNIVERSITÄT BERN

You see a meta-analysis of randomized-controlled trials that compared magnesium with placebo in patients with a heart attack (myocardial infarction). The forest plot for the outcome *death* looks like this:



Review: Intravenous magnesium for acute myocardial infarction
Comparison: 1 Magnesium vs placebo on mortality
Outcome: 3 mortality by dose of magnesium

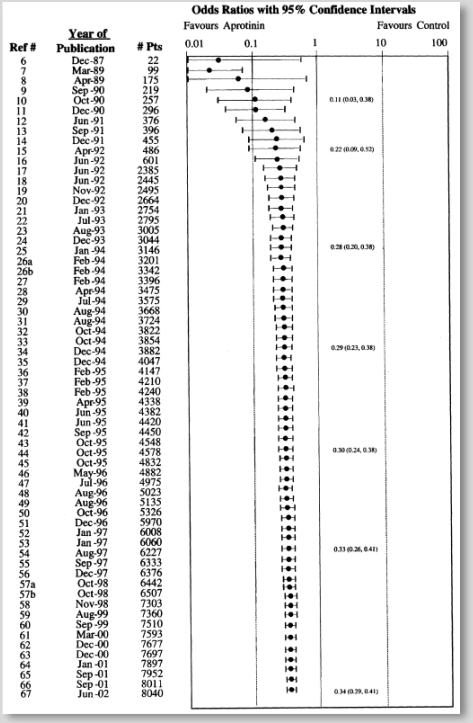
Study or subgroup	Treatment n/N	Control n/N	Odds Ratio M-H,Fixed,95% CI	Weight	Odds Ratio M-H,Fixed,95% CI
1 Mg dose <75 mmol					
Abraham 1987	1/48	1/46		0.0 %	0.96 [0.06, 15.77]
Bhargava 1995	3/40	3/38		0.1 %	0.95 [0.18, 5.00]
Ceremuzynski 1989	1/25	3/23		0.1 %	0.28 [0.03, 2.88]
Gyamlani 2000	2/50	10/50		0.4 %	0.17 [0.03, 0.81]
Nakashima 2004	1/89	3/91		0.1 %	0.33 [0.03, 3.27]
Rasmussen 1986	4/56	14/74		0.4 %	0.33 [0.10, 1.06]
Santoro 2000	0/75	1/75		0.1 %	0.33 [0.01, 8.20]
Shechter 1990	1/50	9/53		0.3 %	0.10 [0.01, 0.82]
Shechter 1991	2/21	4/25		0.1 %	0.55 [0.09, 3.37]
Shechter 1995	4/96	17/98		0.6 %	0.21 [0.07, 0.64]
Singh 1990	6/81	11/81		0.4 %	0.51 [0.18, 1.45]
Smith 1986	2/92	7/93		0.3 %	0.27 [0.06, 1.35]
Thogersen 1995	4/130	8/122		0.3 %	0.45 [0.13, 1.54]
Urek 1996	1/31	0/30		0.0 %	3.00 [0.12, 76.58]
Woods 1992	90/1150	118/1150		4.0 %	0.74 [0.56, 0.99]
Wu 1992	5/125	12/102		0.5 %	0.31 [0.11, 0.92]
Zhu 2002	101/1691	134/1488		4.9 %	0.64 [0.49, 0.84]
Subtotal (95% CI)	**3850**	**3639**		**12.6 %**	**0.59 [0.49, 0.70]**
Total events: 228 (Treatment), 355 (Control)					
Heterogeneity: Chi² = 16.97, df = 16 (P = 0.39); I² =6%					
Test for overall effect: Z = 6.03 (P < 0.00001)					
2 Mg dose >= 75 mmol					
Feldstedt 1991	10/150	8/148		0.3 %	1.25 [0.48, 3.26]
ISIS-4 1995	2216/29011	2103/29039		71.6 %	1.06 [1.00, 1.13]
MAGIC 2000	475/3113	472/3100		14.8 %	1.00 [0.87, 1.15]
Morton 1984	1/40	2/36		0.1 %	0.44 [0.04, 5.02]
Raghu 1999	6/169	18/181		0.6 %	0.33 [0.13, 0.86]
Subtotal (95% CI)	**32483**	**32504**		**87.4 %**	**1.04 [0.99, 1.11]**
Total events: 2708 (Treatment), 2603 (Control)					
Heterogeneity: Chi² = 6.72, df = 4 (P = 0.15); I² =41%					
Test for overall effect: Z = 1.52 (P = 0.13)					
Total (95% CI)	**36333**	**36143**		**100.0 %**	**0.99 [0.94, 1.04]**
Total events: 2936 (Treatment), 2958 (Control)					
Heterogeneity: Chi² = 57.77, df = 21 (P = 0.00003); I² =64%					
Test for overall effect: Z = 0.48 (P = 0.63)					



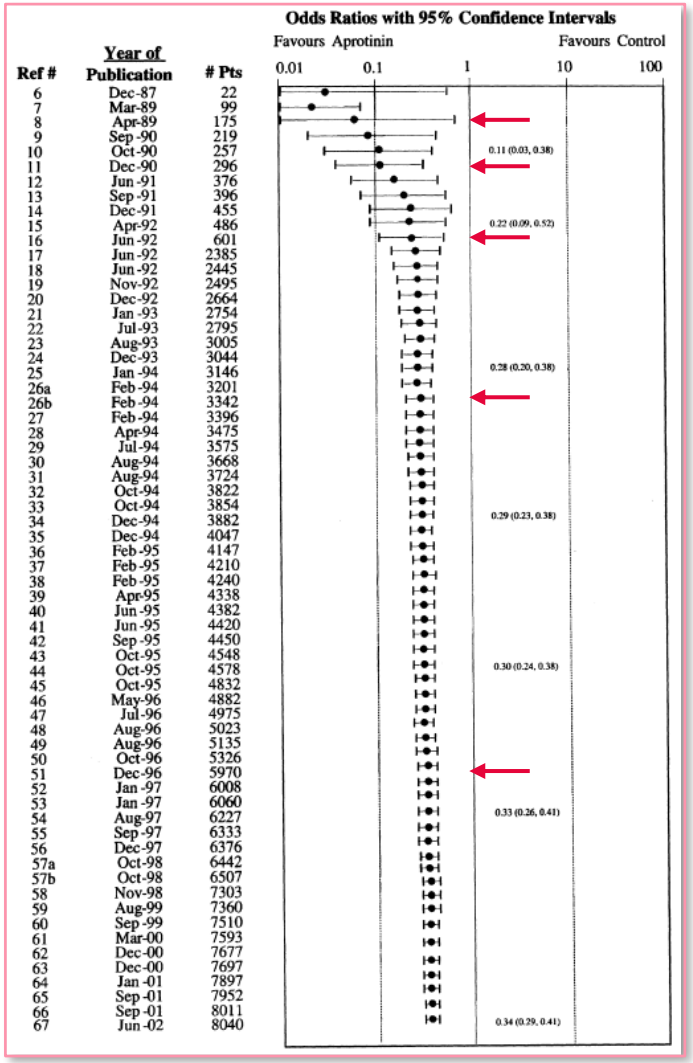
Question #3

Aprotinin

- When do we have sufficient evidence?



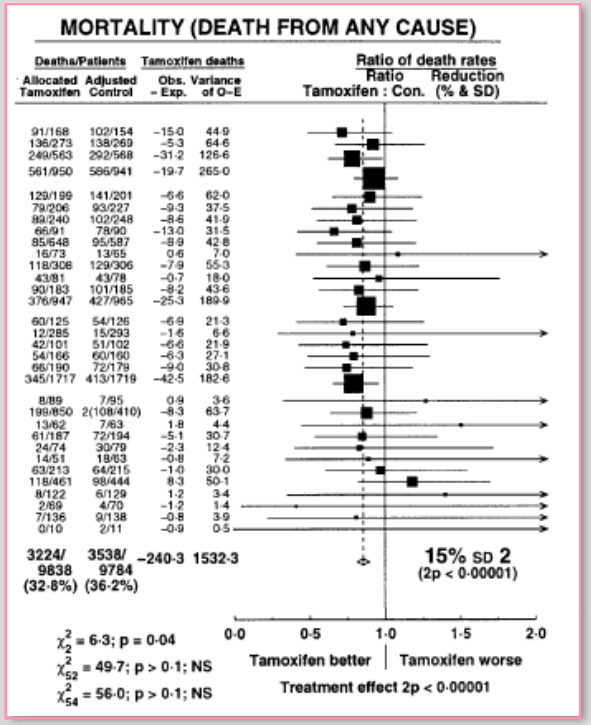
Fergusson 2005


Odds Ratios with 95% Confidence Intervals
Favours Aprotinin
Favours Control
0.01
0.1
1
10
100
Ref # | Year of Publication | # Pts
6 Dec-87 22
7 Mar-89 99
8 Apr-89 175
9 Sep-90 219
10 Oct-90 257
11 Dec-90 296
12 Jun-91 376
13 Sep-91 396
14 Dec-91 455
15 Apr-92 486
16 Jun-92 601
17 Jun-92 2385
18 Jun-92 2445
19 Nov-92 2495
20 Dec-92 2664
21 Jan-93 2754
22 Jul-93 2795
23 Aug-93 3005
24 Dec-93 3044
25 Jan-94 3146
26a Feb-94 3201
26b Feb-94 3342
27 Feb-94 3396
28 Apr-94 3475
29 Jul-94 3575
30 Aug-94 3668
31 Aug-94 3724
32 Oct-94 3822
33 Oct-94 3854
34 Dec-94 3882
35 Dec-94 4047
36 Feb-95 4147
37 Feb-95 4210
38 Feb-95 4240
39 Apr-95 4338
40 Jun-95 4382
41 Jun-95 4420
42 Sep-95 4450
43 Oct-95 4548
44 Oct-95 4578
45 Oct-95 4832
46 May-96 4882
47 Jul-96 4975
48 Aug-96 5023
49 Aug-96 5135
50 Oct-96 5326
51 Dec-96 5970
52 Jan-97 6008
53 Jan-97 6060
54 Aug-97 6227
55 Sep-97 6333
56 Dec-97 6376
57a Oct-98 6442
57b Oct-98 6507
58 Nov-98 7303
59 Aug-99 7360
60 Sep-99 7510
61 Mar-00 7593
62 Dec-00 7677
63 Dec-00 7697
64 Jan-01 7897
65 Sep-01 7952
66 Sep-01 8011
67 Jun-02 8040
0.11 (0.03, 0.38)
0.22 (0.09, 0.52)
0.28 (0.20, 0.38)
0.29 (0.23, 0.38)
0.30 (0.24, 0.38)
0.33 (0.26, 0.41)
0.34 (0.29, 0.41)

Question #4

Tamoxifen

– Several homogeneous non-significant trials



Early Breast Cancer Trialists' Collaboration Group 1998

Clinical research

Anything to do with us?

1.B SCREENING VISIT: Medical history

To be completed by the study staff. Please use CAPITAL LETTERS

Site: Subject ID:BOOST - __ __ -__ __ -__ __ __ Visit date: (dd-mm-yyyy)

Current Medications

Include over-the-counter preparations (OTC), dietary and vitamin supplements , herbal supplements, homeopathic agents and painkillers

Medication	Dose and units	Frequency	Route	Indication	Start date	Stop date

*** Females only:**

Use of any contraceptive method: Y N Contraceptive method discussed with the doctor Y N

Please specify contraceptive method: _____

If you require more space please use additional CRF page EU-COVAT-2BOOSTAVAC_CRFv1.0 04.03.22

Voting

To think about

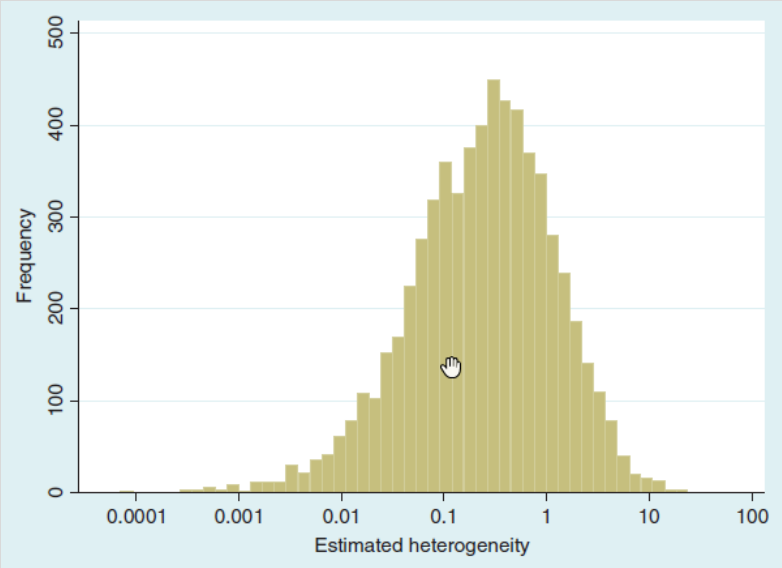
UNIVERSITÄT BERN

- Is reproducibility the right concept?
- Is knowledge from clinical trials (*evidence* (the *truth*?)) rather cumulative?
- Are most published research studies false? → Maybe
- Do we have a reproducibility crisis in clinical research? → Maybe not

Crisis?

UNIVERSITÄT BERN

- 14,886 meta-analyses with 77,237 individual trials
- 57% of meta-analyses had no statistical between-trial heterogeneity but 43% had →



Turner 2012

Voting

To think about

UNIVERSITÄT BERN

- Is reproducibility the right concept?
- Is knowledge from clinical trials (*evidence* (the *truth*?)) rather cumulative?
- Are most published research studies false (Ioannidis 2005)? → Maybe
- Do we have a reproducibility crisis in clinical research? → Maybe not
- Quantification versus testing

Overarching objectives of a trial

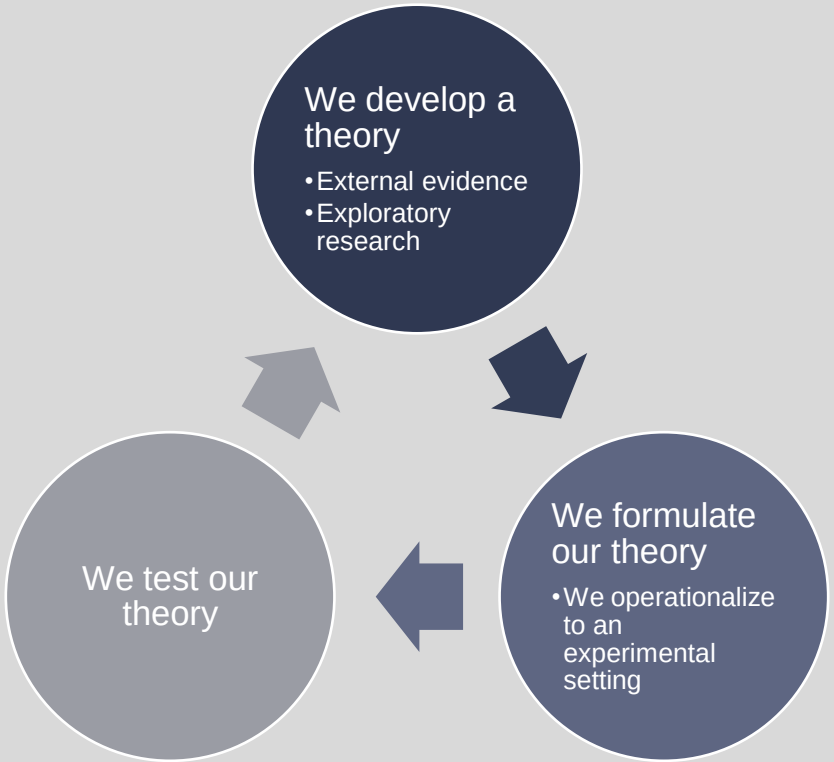
UNIVERSITÄT BERN

- Experiment to quantify cause-and-effect i.e. exposure/intervention → outcome

- Mechanistic (scientific research)
- (Clinical) Practice (evaluative research)
  - Commercial/industry: to sell a product (e.g. pharmaceutical, device, …) to make money
  - Academic: to change practice, make a career, …
  - Mandated (UK NIH): to resolve uncertainty and optimize health care provision

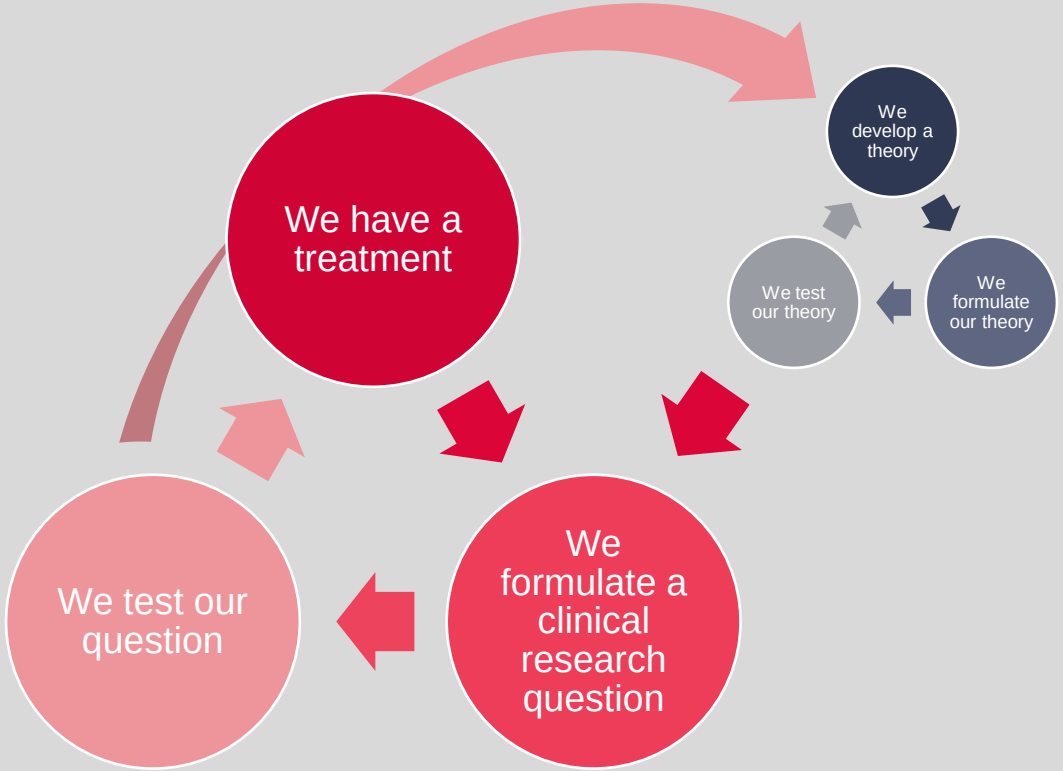
Objective of (mechanistic) experimenting

Discovery (via proofs)



Objective of (clinical) experimenting

Evaluation/proofing



Other reasons that are discussed

Explanations that might be of relevance for clinical research, too

- Sampling variability (chance)
- Differences across studies that act as effect modifiers/moderators
- How do we measure (quantify) replicability?
  - Replicating statistical significance is probably not a criterion that is affordable (van Zwet 2022)
- Fraud and misconduct

Traceability

and what we do about them in clinical research

UNIVERSITÄT BERN

- Documentation …
- Trial protocol, case report forms, research databases (user management and audit trail), …
- The concept of the Source and Source Data Location Log
  - Source (proof that data* exists) → Case Report Form → Database → Statistical analysis
  - Source Data Location Log defining where original data can be found
    - Important: if multiple potential sources exists (as is usual in clinical medicine/health care) → hierarchy of sources!

* A (set of) values, information

Questionable research practices

and what we do about them in clinical research

UNIVERSITÄT BERN

- Publication bias
- Outcome reporting bias
- Discrepant reporting

→Trial (and results) registration

→(Registered reports (previous Lancet initiative (not active anymore …), getting increased awareness in social sciences)

- Low power

→Sample size calculation …

Analysing data and interpreting results

and what we do about them in clinical research

UNIVERSITÄT BERN

- P-hacking
  - Fishing for significant resuls
- **H**ypothesiz**ing** **A**fter **R**esults are **K**nown (Kerr 1998)
  - post hoc hypothesis in the introduction of a research report as if it were an a priori hypothesis
- Researchers degree of freedom (Simmons 2011)
- Garden of Forking Paths (Gelman 2013)
  - Increase in false positive results even without questionable research practices
  - Multiple choices and many correct analytical approaches

→ Statistical Analysis Plan before (the majority) of participants is enrolled and looking into the data

Outlook

(Data) Sharing and transparency

To ensure trust (and further research)

- Trial protocol
- Patient information and consent form
- Data Management Plan
- Statistical Analysis Plan
- Monitoring reports
- Documentation on protocol deviations
- (Narratives?)
- Data dictionary
- Data
- Statistical code
- …

Thank you

for your attention!

Sven Trelle, CTU Bern

April 27, 2022

Lots of ideas/inspiration stolen from Steven Goodman, Sander Greenland, Andrew Gelman, David Spiegelhalter, …

References

- Baker 2016. Nature;533:452.
- Camerer 2018. Nat Hum Behav;2:637.
- Dahabreh 2019. Eur J Epidemiol;34:719.
- Early Breast Cancer Trialists' Collaboration Group 1998. Lancet;352;403.
- Errington 2021. eLife 2021;10:e71601
- Fergusson 2005: Clin Trials;2:218.
- Freedman 2015. PLoS Biol;13:e1002165.
- Gelman 2013. https://osf.io/n3axs/#!
- Goodman 2016. Sci Transl Med;8:1.
- Ioannidis 2005. JAMA;294:218.
- Ioannidis 2005. PLoS Med. 2:e124.
- Kerr 1998. Pers Soc Psychol Rev;2:196.
- Open Science Collaboration 2015. Science;349:943.
- Reichenbach 2007. Ann Intern Med;17:580.
- Simmons 2011. Psychol Sci;22:1359.
- Smaldino 2016. Royal Soc Open Sci;3:160384
- Spiegelhalter 2019. The art of statistics. London: Pelican.
- Turner 2012. Int J Epidemiol;41:818.
- Van Zwet 2022. Stat Med. Epub ahead of print.