Cochrane

Do Cochrane and non-Cochrane editors & authors prefer reporting statements based on statistically significant differences or do they prefer non-binary options?

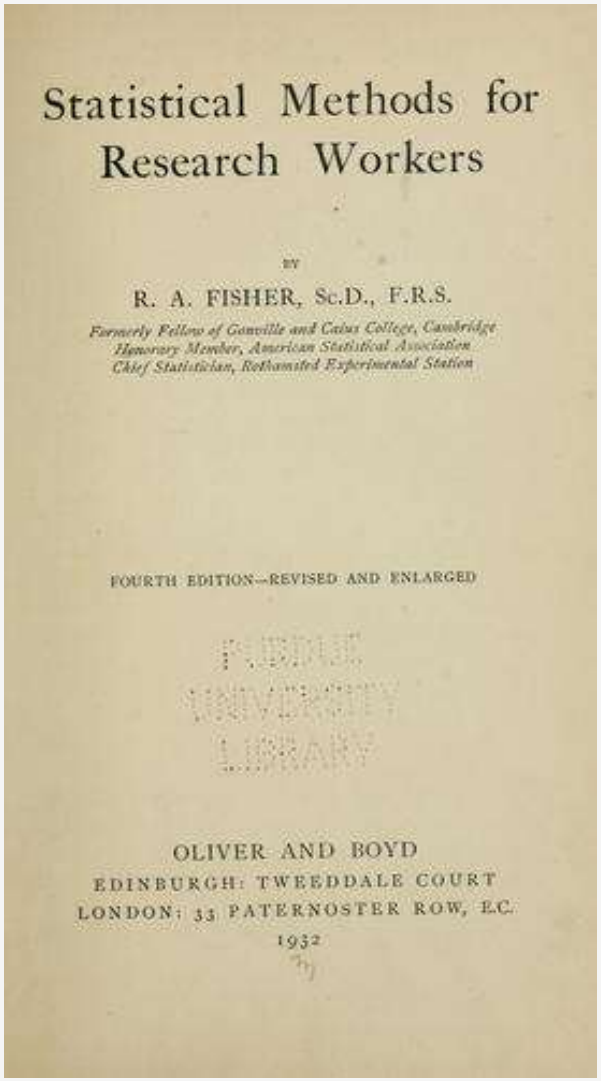
Ciapponi A, Glujovsky D, Bardach A. ▷ No conflict of interest
aciapponi@iecs.org.ar

Trusted evidence.
Informed decisions.
Better health.

Cochrane Argentina

IECS INSTITUTO DE EFECTIVIDAD CLINICA Y SANITARIA

cegyr



Statistical Methods for Research Workers

BY

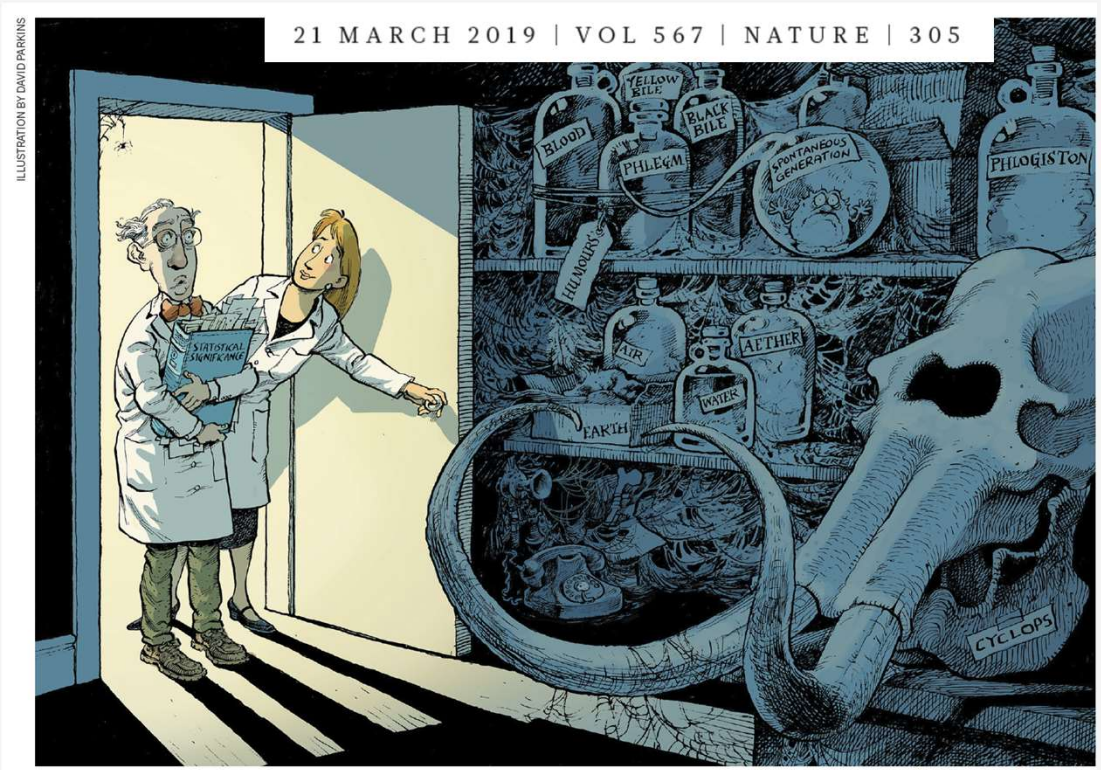
R. A. FISHER, Sc.D., F.R.S.

Formerly Fellow of Gonville and Caius College, Cambridge
Honorary Member, American Statistical Association
Chief Statistician, Rothamsted Experimental Station

FOURTH EDITION—REVISED AND ENLARGED

OLIVER AND BOYD
EDINBURGH: TWEEDDALE COURT
LONDON: 33 PATERNOSTER ROW, E.C.
1932

P



Retire statistical significance

Valentin Amrhein, **Sander Greenland**, **Blake McShane** and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

EDITORIAL **Open Access**

There is life beyond the statistical significance

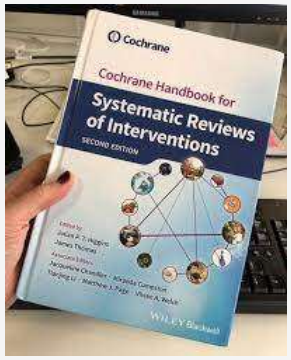
Ciapponi *et al. Reprod Health* *(2021) 18:80*
https://doi.org/10.1186/s12978-021-01131-w

Check for updates

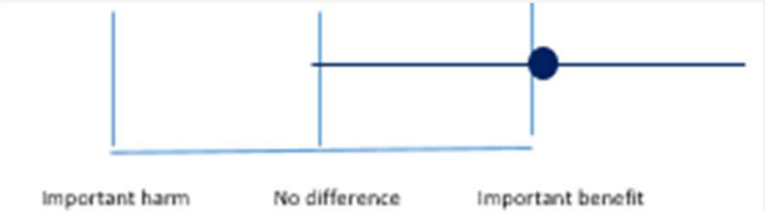
BMC

- For decades the p value-based interpretation and reporting of results dominated the publications, but scientific community agrees that this binary approach is not enough and suggested a systemic reform to change this paradigm.

- The Cochrane Handbook, recommends reporting the point estimate, the CI + exact P-value, MIDs, some narrative statements, and against binary approaches:



> Review authors should not describe results as 'statistically significant', 'not statistically significant' or 'non-significant' or unduly rely on thresholds for P values, but report the confidence interval together with the exact P value. Chapter 15



- **Which is the approach of Cochrane and non-Cochrane editors and authors for interpretation and reporting this case?**

Cochrane

NEW SECTION

Methods

Cochrane

Stakeholders surveyed

- **Cochrane editors** (N=65)
- **Cochrane authors,** that published reviews from 1/1/23 to 7/25/23 (N=321)

Source: Archie

- **Non-Cochrane editors** (N=20)
- **Non-Cochrane authors** (N=322)

Source: the 20 highest impact factor in the "General Medicine" and "Internal Medicine" categories in 2021 (edition of Clarivate Analytics Journal Citation Report) with available e-mail

Cochrane

Stakeholders had to choose the binary or non-binary option that better expresses the results for the following scenario:

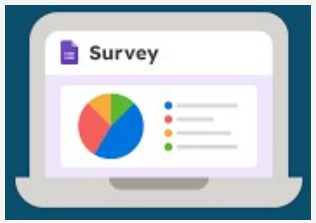
¨After exhaustive literature searches, a systematic review identified only two pivotal RCTs that evaluated the mortality of drug X versus placebo (P) in patients with a rare genetic disease. The risk of bias for all domains was low in both RCTs (assessed using the Cochrane RoB-2 tool), and there was no methodological, clinical, or statistical heterogeneity between studies. The meta-analysis showed the following results:¨

	X drug	Placebo
Mortality risk	**26%** (10/39)	**45%** (18/40)
Risk difference	**With X 19% lower mortality** (95% CI **40% lower** to **1% higher**)	
Risk Ratio	**0.57** (95% CI **0.30** a **1.08**)	
P value	**0.0721**	

Clinical important difference with CI crossing the null effect

Cochrane

Please, select only the statement that better reflects the interpretation of the results, even if more than one is correct:



1. Mortality with **X** **is lower** than with **Placebo (P)**
2. Mortality with **X** **is probably lower** than **P**, **but no statistically significant** differences were found
3. Mortality with **X** **is probably lower** than with **P**, **but** the probability that the difference is **due to chance is 7%**
4. Mortality with **X** is **possibly lower** than with **P**, but the possibility that the difference is due to **chance is 7%** (this one is similar to statement 3, but replacing probably by possibly)
5. Mortality with **X** is **probably lower** than with **P**, **but** the **confidence interval (CI)** is **compatible with** both a **reduction** and an **increase** in mortality.
6. Mortality with **X** is **possibly lower** than with **P**, but the confidence interval is compatible with both a **reduction** and an **increase** in mortality (this one is similar to statement 5, but replacing probably by possibly)
7. **No differences were found** between **X** and **P**
8. Intervention **X** **did not show higher mortality** than **P**
9. **No statistically significant differences were found** between **X** and **P**

Could you, please, justify your answer?

Cochrane

GRADEpro GDT

The case

Help

Should drug X vs. placebo be used for rare condition Y?

Bottom panel | Explanations

Drug X compared to placebo for rare condition Y

Certainty assessment							Summary of findings					
							№ of patients		Effect			
№ of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Drug X	Placebo	Relative (95% CI)	Absolute (95% CI)	Certainty	Importance
Mortality (follow-up: mean 12 months)												
2	randomised trials	not serious	not serious	not serious	?	none	10/39 (25.6%)	18/40 (45.0%)	**RR 0.57** (0.30 to 1.08)	**194 fewer per 1,000** (from 315 fewer to 36 more)	-	CRITICAL

Imprecision

- not serious
- serious
- very serious
- extremely serious

Cochrane

GRADE

Updates on rating imprecision

ELSEVIER
Journal of Clinical Epidemiology 150 (2022) 216–224
Journal of Clinical Epidemiology

GRADE GUIDANCE SERIES

GRADE Guidance 34: update on rating imprecision using a minimally contextualized approach

Linan Zeng[a,b,*], Romina Brignardello-Petersen[b], Monica Hultcrantz[c], Reem A. Mustafa[d], Mohammad H. Murad[e], Alfonso Iorio[b,f], Gregory Traversy[g], Elie A. Akl[h], Martin Mayer[i,j,k], Holger J. Schünemann[b,f], Gordon H. Guyatt[b,f]

ELSEVIER
Journal of Clinical Epidemiology 150 (2022) 225–242
Journal of Clinical Epidemiology

GRADE GUIDANCE SERIES

GRADE guidance 35: update on rating imprecision for assessing contextualized certainty of evidence and making decisions

Holger J. Schünemann[a,b,c,q,*], Ignacio Neumann[b,d], Monica Hultcrantz[e], Romina Brignardello-Petersen[b], Linan Zeng[b,f], M Hassan Murad[g], Ariel Izcovich[h], Gian Paolo Morgano[b], Tejan Baldeh[b], Nancy Santesso[a,b], Carlos Garcia Cuello[a,b], Lawrence Mbuagbaw[a,b], Gordon Guyatt[b,c], Wojtek Wiercioch[a,b], Thomas Piggott[a,b], Hans De Beer[i], Marco Vinceti[j], Alexander G. Mathioudakis[k], Martin G. Mayer[l,m,n], Reem Mustafa[o], Tommaso Filippini[j], Alfonso Iorio[b,c], Robby Nieuwlaat[a,b], Maura Marcucci[b,c], Pablo Alonso Coello[p], Stefanos Bonovas[q,r], Daniele Piovani[q,r], George Tomlinson[s,t], Elie A. Akl[b,u], for the GRADE Working Group

ELSEVIER
Journal of Clinical Epidemiology 147 (2022) 69–75
Journal of Clinical Epidemiology

Other GRADE Papers

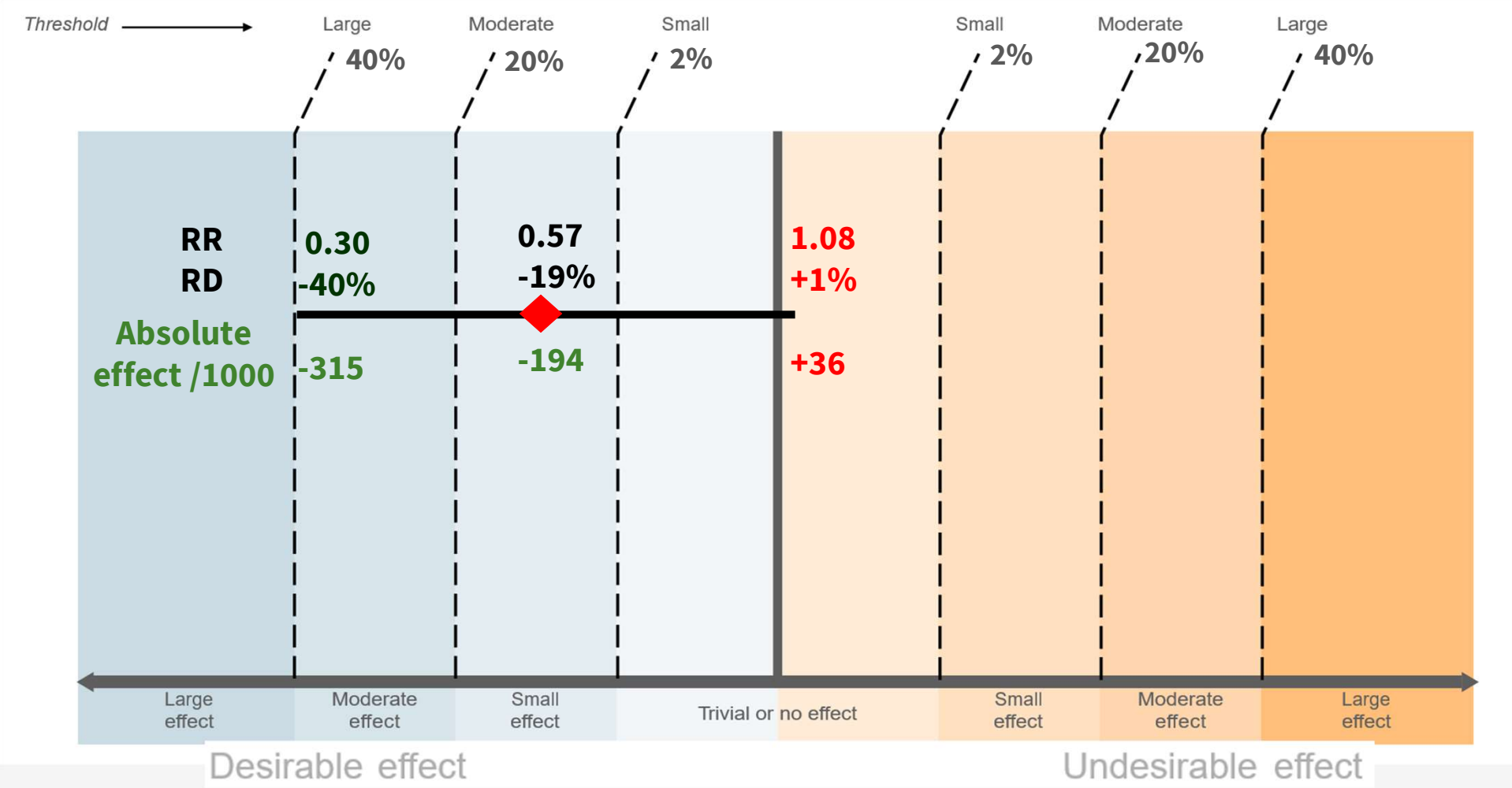
Using Explicit Thresholds were valuable for judging Benefits and Harms in partially contextualized GRADE Guidelines

Ignacio Neumann[a,b,c,*], Eduardo Quiñelen[b], Paula Nahuelhual[b], Pamela Burdiles[b], Natalia Celedón[b], Katherine Cerda[b], Paloma Herrera-Omegna[b], Patricia Kraemer[b], Karen Dominguez Cancino[b,c,d], Juan Pablo Valenzuela[b], Dino Sepúlveda[b], Gian Paolo Morgano[c], Elie A. Akl[c,e], Holger J. Schünemann[c]

Cochrane

GRADE

Thresholds and ranges for trivial, small, moderate and large effects



Calculating the review informatin size (RIS) when effects are large	Prop	Prop (%)
User-entered data		
Threshold for small effect (MID)	0.02	2%
Threshold for moderate effect	0.2	20%
Threshold for large effect	0.4	40%
Baseline risk	0.45	45%
Type I error (alpha)	0.05	5%
Power	0.8	80%
Review size (total number of participants)	800	

Results of RIS calculation

Threshold	№ of participants (studies)	Sample size to rule out	GRADE a
Small (MID)	79 (2 RCTs)	958	Rate down
Moderate effect		958	Rate down
Large effect		94	Do not rate down

e

Should drug X vs. placebo be used for rare condition Y?

Bottom panel Explanations

Drug X compared to placebo for rare condition Y

Certainty assessment							Summary of findings					
№ of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	№ of patients		Effect		Certainty	Importance
							Drug X	Placebo	Relative (95% CI)	Absolute (95% CI)		
Mortality (follow-up: mean 12 months)												
2	randomised trials	not serious	not serious	not serious	very serious [a]	none	10/39 (25.6%)	18/40 (45.0%)	**RR 0.57** (0.30 to 1.08)	**194 fewer per 1,000** (from 315 fewer to 36 more)	⊕⊕◯◯ Low	CRITICAL

a Downgragraded to levels because the confidence interval crossed two effect thresholds

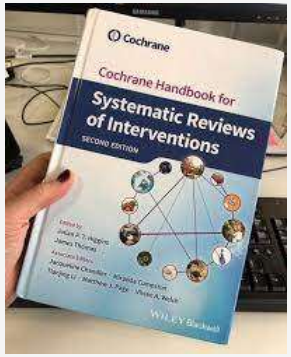


Table 1 Suggested narrative statements for phrasing conclusions

Certainty of the evidence	Effect size	Suggested statements for conclusions (replace X with intervention, choose 'reduce' or 'increase' depending on the direction of the effect, replace 'outcome' with name of outcome, include 'when compared with Y' when needed)
⊕⊕◯◯ Low	Moderate	X may reduce/increase outcome The evidence suggests X reduces/increases outcome X may result in a reduction/increase in outcome The evidence suggests X results in a reduction/increase in outcome

~~Probable~~ **Possible**

Statement assessment (not exclude the reporting of numbers)

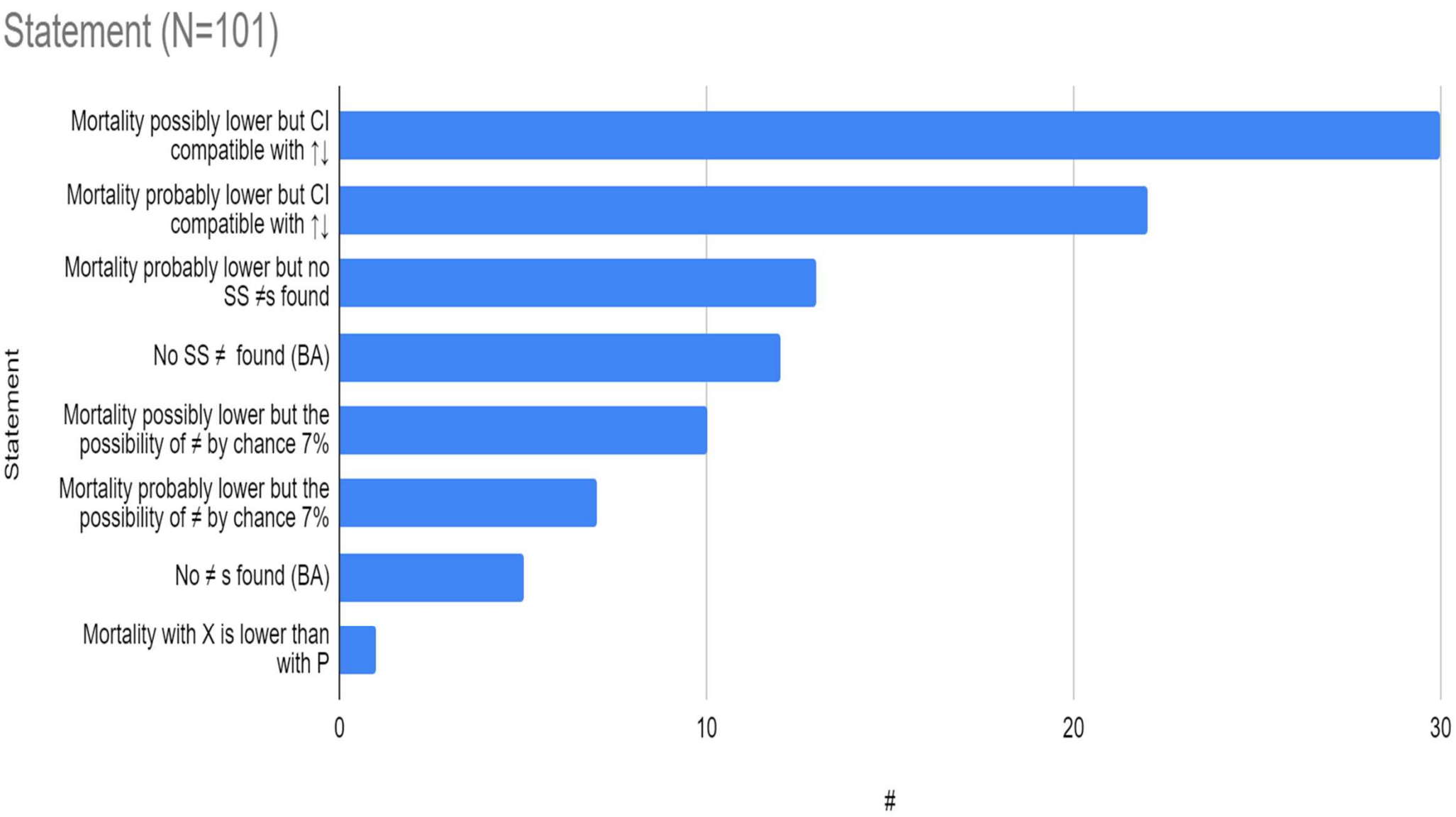
1. Mortality with X **is lower** than with P (**Binary approach**)
2. Mortality with X **is probably lower** than P, **but no statistically significant** differences were found
3. Mortality with X **is probably lower** than with P, **but** the probability that the difference is **due to chance is 7%**
4. Mortality with X is **possibly lower** than with P, but the possibility that the difference is due to **chance is 7%**
5. Mortality with X is **probably lower** than with P, **but** the **confidence interval (CI)** is **compatible with** both a **reduction** and an **increase** in mortality.
6. Mortality with X is **possibly lower** than with P, but the CI is compatible with both a **reduction** and an **increase** in mortality
7. **No differences were found** between X and P (**Binary approach** not high CoE: serious imprecision)
8. Intervention X **did not show higher mortality** than P (**Binary approach** not high CoE: serious imprecision)
9. **No statistically significant differences were found** between X and P (**Binary approach** not high CoE: serious imprecision)

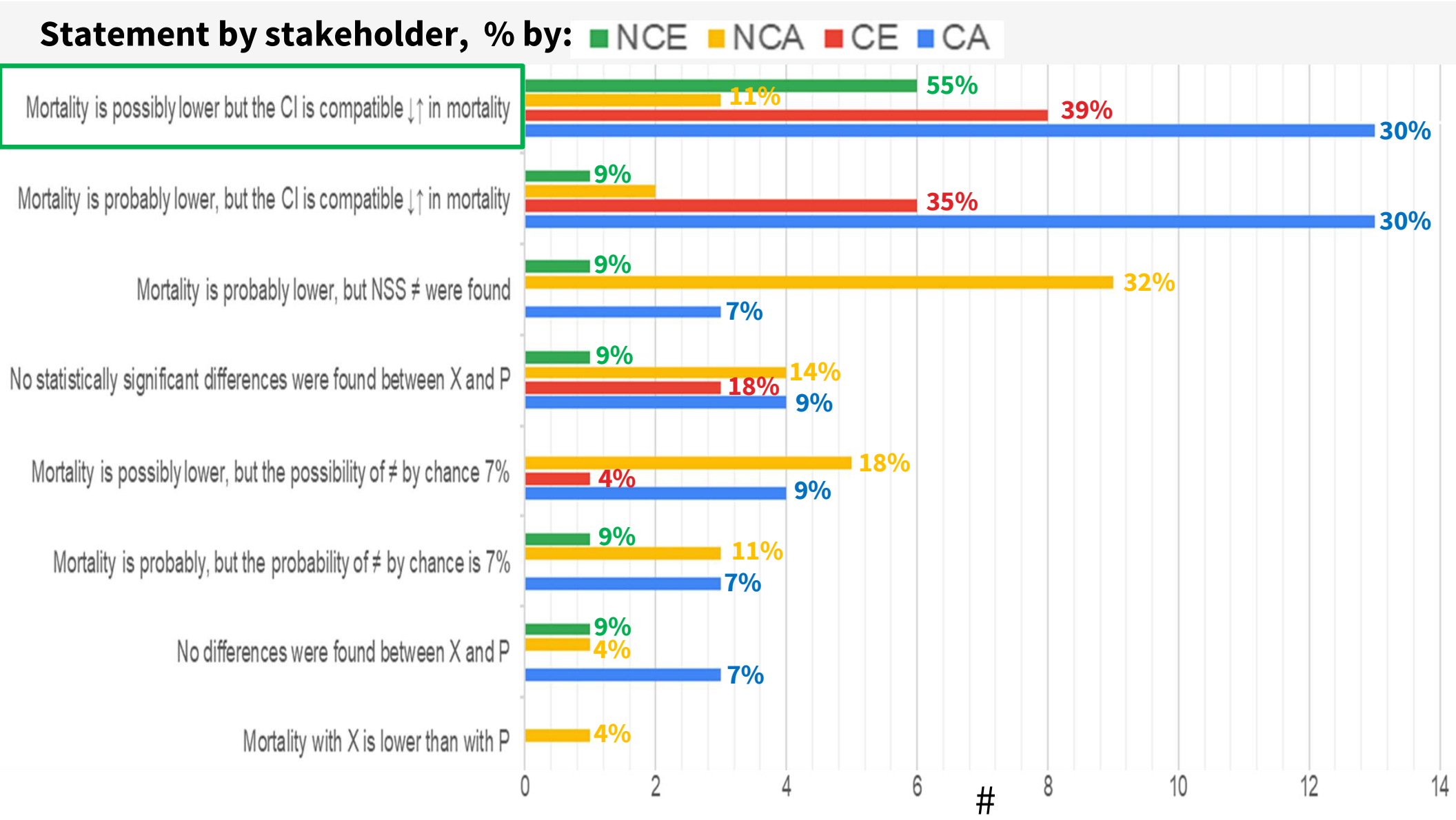
Cochrane

NEW SECTION

Results

Cochrane


Statement (N=101)
Mortality possibly lower but CI compatible with ↑↓
Mortality probably lower but CI compatible with ↑↓
Mortality probably lower but no SS ≠s found
No SS ≠ found (BA)
Mortality possibly lower but the possibility of ≠ by chance 7%
Mortality probably lower but the possibility of ≠ by chance 7%
No ≠ s found (BA)
Mortality with X is lower than with P
Statement
0
10
20
30
#


Statement by stakeholder, % by:
NCE
NCA
CE
CA
Mortality is possibly lower but the CI is compatible ↓↑ in mortality
55%
11%
39%
30%
Mortality is probably lower, but the CI is compatible ↓↑ in mortality
9%
35%
30%
Mortality is probably lower, but NSS ≠ were found
9%
32%
7%
No statistically significant differences were found between X and P
9%
14%
18%
9%
Mortality is possibly lower, but the possibility of ≠ by chance 7%
18%
4%
9%
Mortality is probably, but the probability of ≠ by chance is 7%
9%
11%
7%
No differences were found between X and P
9%
4%
7%
Mortality with X is lower than with P
4%
0
2
4
6
8
10
12
14
#

Cochrane Authors (n=43, 13%)

↓ 2 levels GRADE because the OIS is not reached (<100 events, and CI overlapping the line of no effect). The best "mortality with X may be lower than with placebo but...

the CI is wide, so benefit on mortality cannot be completely excluded

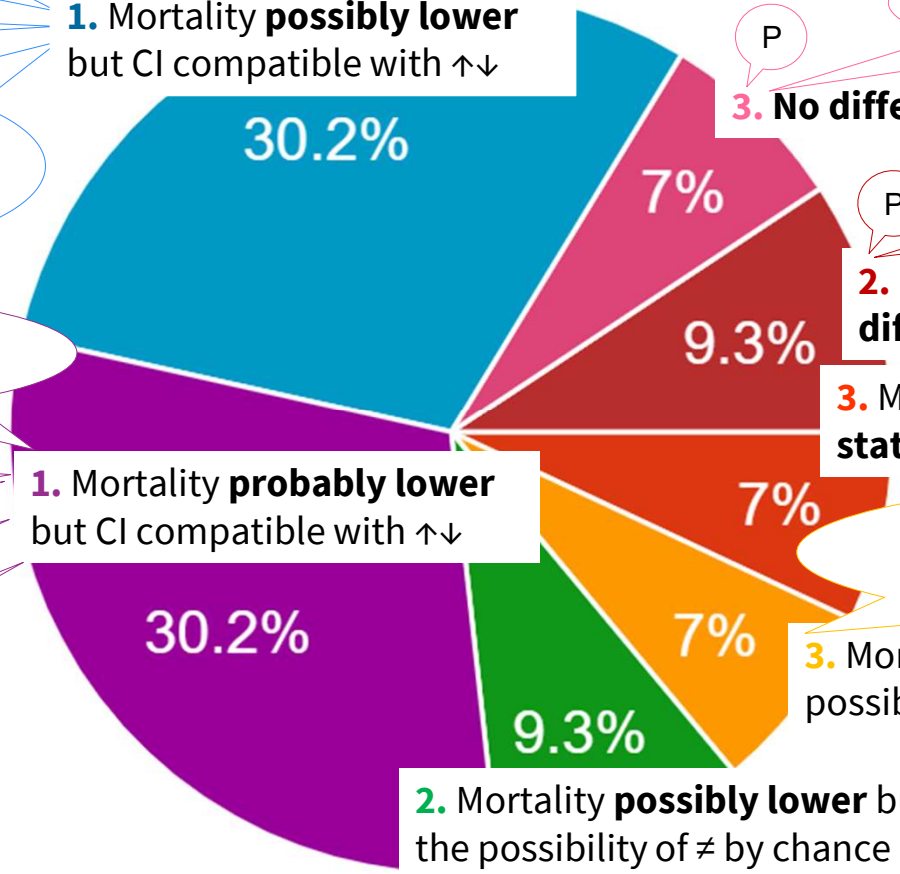
Low certainty, NSS, Wide CI

The # of participants and events so low that I'd rather say possibly than probably

there is probably/may be little or no ≠ between groups

P

P



1. Mortality **possibly lower** but CI compatible with ↑↓

3. No differences found (BA)

CIs include the null effect probably due to low power

2. No statistically significant differences found (BA)

3. Mortality **probably lower but no statistically significant** differences found

I prefer NNTB 5 (RD -19%). We should not destroy the chance of the patients to receive these medication only because an arbitrary threshold for SS

3. Mortality **probably lower** but the possibility of ≠ by chance 7%

Mortality with X might be either ≤ than with P (unlikely to be higher)

2. Mortality **possibly lower** but the possibility of ≠ by chance 7%

Absolute effects are as important as relative effects

CI compatible with ↓ & ↑ in mortality

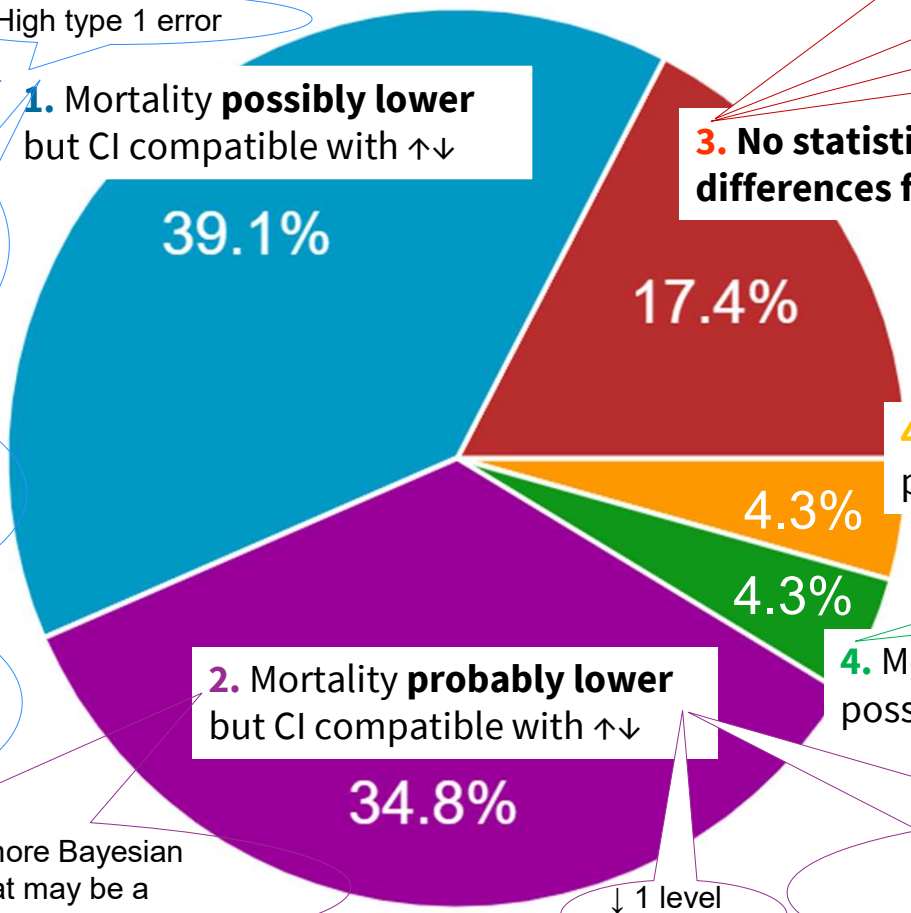
Add p value

Moderate certainty

CI including null effect => No ≠

1. Mortality **probably lower** but CI compatible with ↑↓

Cochrane Editors (n=23, RR 35%)



Only 79 participants, wide CI, ↓ # events

High type 1 error

If the studies were assessed according to the trustworthiness screening tool or the Research Integrity Assessment tool or similar

↓ 2 levels GRADE guide 34 update on rating imprecision using a minimally contextualized approach

+ explicitly statement that there were NSS ≠ & taking account other outcomes to assist with the interpretation

The probably is driven by a more Bayesian view…but is also possible that may be a small increase in mortality

↓ 1 level GRADE

Probably instead of possibly because of the very large possible benefit vs the very small possible harm (1%↑)

I would prefer "uncertain" in here or "little or no difference"

Mortality is lower with X, although the evidence is not strong enough. It is more neutral.

I would prefer insufficient evidence of a ≠

Non-Cochrane Editors (n=11, RR 55%)

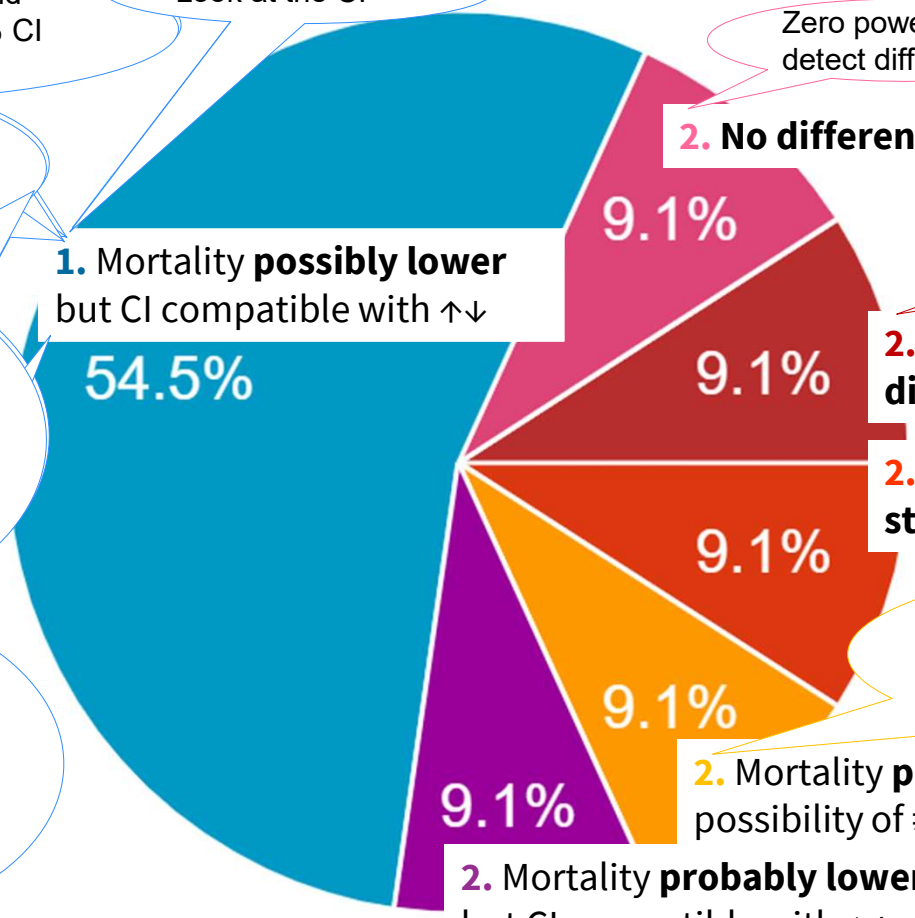
The point estimates for RD and RR favor Drug X, but the 95% CI indicate lack SS

Look at the CI

I prefer a statement that says we do not know and just gives the raw numbers

I don't think that NSS≠ tells the whole story. his result is SO close to significance, with an effect size that is substantial, that I can't ignore the possibility that the benefit is real

Due to the very small cumulative sample size, there is a substantial possibility that a future study could change the direction of the



1. Mortality **possibly lower** but CI compatible with ↑↓

Zero power to detect differences

2. **No differences found (BA)**

Results NS given the small numbers

2. **No statistically significant differences found (BA)**

2. Mortality **probably lower but no statistically significant** differences found

I don't think we should use strictly frequentist interpretations of p-values and CI, but rather try to interpret the data as best we can and allowing for uncertainty

2. Mortality **probably lower** but the possibility of ≠ by chance 7%

2. Mortality **probably lower** but CI compatible with ↑↓

Cochrane

Non-Cochrane authors (n=28, RR 9%)

GRADE guide 34

↓↓ NNTB but low power

5. Mortality **probably lower** but CI compatible with ↑↓

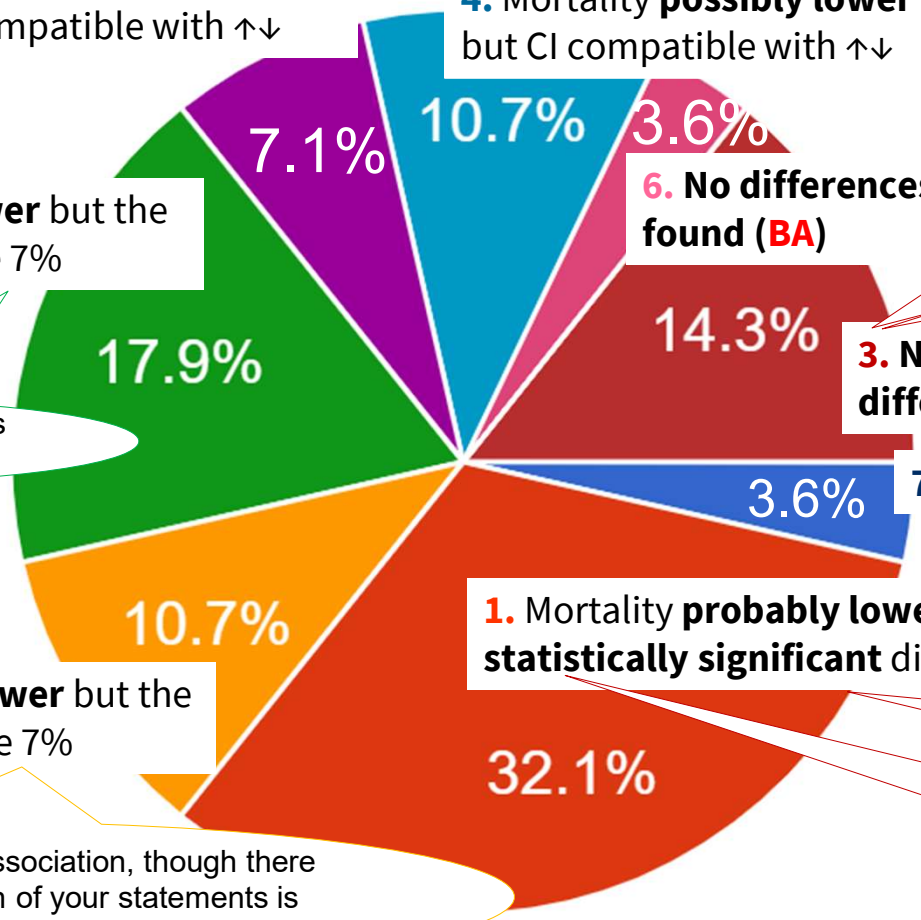
4. Mortality **possibly lower** but CI compatible with ↑↓

The 2 studies were described as being of sound methodology. It is not justified to imply that there are differences when the available high-quality data does not support this conclusion.

2. Mortality **possibly lower** but the possibility of ≠ by chance 7%

GRADE guide 34

The biggest concern here is the very small sample size



6. No differences found (BA)

P

3. No statistically significant differences found (BA)

7. Mortality **is lower** with X

Some would say this shows a trend, but one would need more data

1. Mortality **probably lower but no statistically significant** differences found

4. Mortality **probably lower** but the possibility of ≠ by chance 7%

CIs include the null effect probably due to low power

CI crosses 1; > 0.05. The outcome is tremendously important justifying clinical action

This likely represents a true association, though there is not power to conclude which of your statements is most accurate

NEW SECTION

Conclusions

Cochrane

- There is high heterogeneity of selected statements
- 1/5 Cochrane and non-Cochrane editors & authors still select binary approaches
- The GRADE approach is not always considered to define the certainty of evidence.
- A very low proportion (3%) explicitly considered the GRADE update for rating imprecision (not at all among non-Cochrane editors)
- Probable the best option: *"Mortality may be lower with X than with P, but the CI is compatible with both a reduction and an increase in mortality"*
- Including the probability of chance in the statement is better than only referring to the statistical significance, but it could be also informed by adding the p-value to other effect measures in numbers.

- The case of clinical important difference with CIs crossing the null effect is still an reporting and interpretation challenge.
- The moderate response rate does not warrant representativeness, but suggests that not responders could have a worse performance.
- There are several correct reporting statements and it would be desirable a higher consistency.
- The GRADE update should be strongly diffused.
- Further research should assess the interpretability of the reporting statements.

Thank you!

Cochrane Argentina