A Fresh Look at the Two-Study Paradigm

Leonhard Held

University of Zurich UZH

EFSPI Scientific Meeting: Reproducibility in Clinical Research

Bristol-Myers Squibb Belgium SA

Braine-l'Alleud

November 22, 2019

Introduction

- Replicability of research findings is crucial to the credibility of science.
- Large-scale replication projects have been conducted in the last years.
- Such efforts help to assess to what extent results from original studies can be confirmed in independent replication studies.

UZH

Center for Reproducible Science

The Two-Trials Rule

- FDA/EMA requires

  > *"at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness"*

  for many diseases.
- Usually implemented requiring one-sided $p \leq \alpha = 0.025$ in two independent studies.

The Two-Trials Rule

- FDA/EMA requires

  > *“at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness”*

  for many diseases.
- Usually implemented requiring one-sided $p \leq \alpha = 0.025$ in two independent studies.
- However, this may not reflect the available evidence:
  - $p_1 = p_2 = 0.024$ leads to claim of success.
  - $p_1 = 0.026$ and $p_2 = 0.001$ does not lead to claim of success.

The Two-Trials Rule

- FDA/EMA requires

  > *“at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness”*

  for many diseases.
- Usually implemented requiring one-sided $p \leq \alpha = 0.025$ in two independent studies.
- However, this may not reflect the available evidence:
  - $p_1 = p_2 = 0.024$ leads to claim of success.
  - $p_1 = 0.026$ and $p_2 = 0.001$ does not lead to claim of success.
- It is also not clear how to extend the rule to results from $n > 2$ studies:
  - Requiring at least 2 out of n studies to be significant is too lax.
  - Requiring all n studies to be significant is too stringent.

Combining and Pooling *P*-Values

- Fisher's combined method is sometimes used, but also has problems:
  - $p_1 = 0.0001$ and $p_2 = 0.5$ gives Fisher's $p = 0.0005 < 0.025^2$.
  - $p_1 = 0.01$ and $p_2 = 0.01$ gives Fisher's $p = 0.001 > 0.025^2$.

Combining and Pooling *P*-Values

- Fisher's combined method is sometimes used, but also has problems:
  - $p_1 = 0.0001$ and $p_2 = 0.5$ gives Fisher's $p = 0.0005 < 0.025^2$.
  - $p_1 = 0.01$ and $p_2 = 0.01$ gives Fisher's $p = 0.001 > 0.025^2$.
- Similar problems for Stouffer's pooled method based on (weighted) average of Z-scores (meta-analysis).

Combining and Pooling *P*-Values

- Fisher's combined method is sometimes used, but also has problems:
  - $p_1 = 0.0001$ and $p_2 = 0.5$ gives Fisher's $p = 0.0005 < 0.025^2$.
  - $p_1 = 0.01$ and $p_2 = 0.01$ gives Fisher's $p = 0.001 > 0.025^2$.
- Similar problems for Stouffer's pooled method based on (weighted) average of Z-scores (meta-analysis).
- Combinations with the two-trials rule have been proposed in Rosenkrantz (2002) and Maca *et al.* (2002), but require specification of a relaxed criterion α' for significance of the two individual trials.

Variations on the Two-Trials Rule

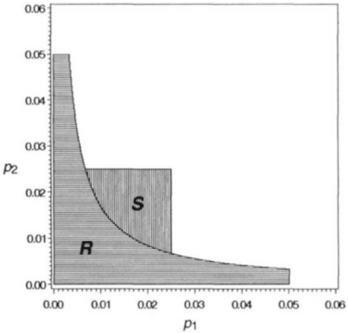
Restrictions on study-specific p-values

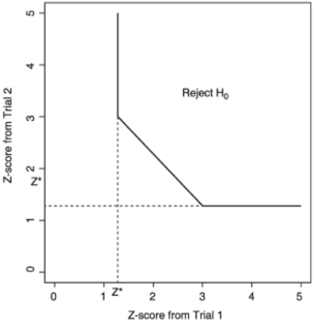
Rosenkranz (2002)

$\alpha' = 0.05$

Maca *et al.* (2002)

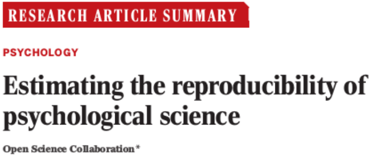
$\alpha' = 0.1$





The Reproducibility of Psychological Science

Open Science Collaboration (2015, *Science*)

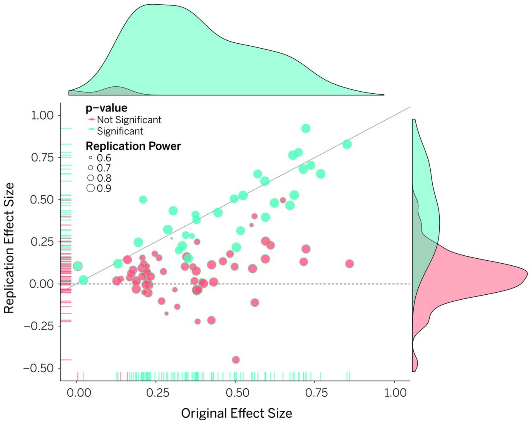


RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

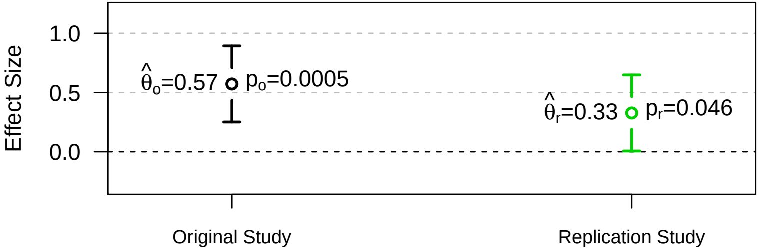
Estimating the reproducibility of psychological science

Open Science Collaboration*



Analysis of Replication Studies

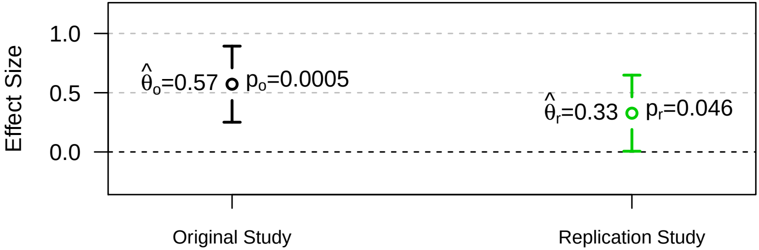
Effect estimates with 95% confidence interval



$\hat{\theta}_o$	Effect estimate	$\hat{\theta}_r$
σ_o	Standard error	σ_r
n_o	Sample size	n_r
p_o	*p*-value	p_r

Replication Success

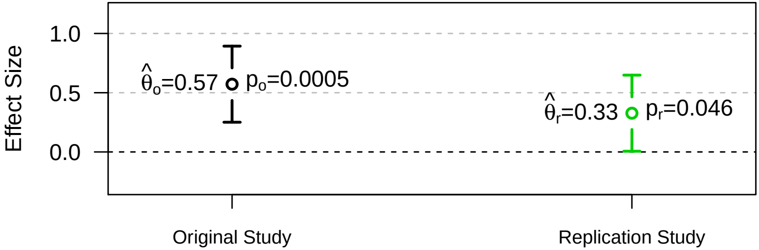
Lack of a single accepted definition



1. Assessment of significance (as in the two-trials rule)

Replication Success

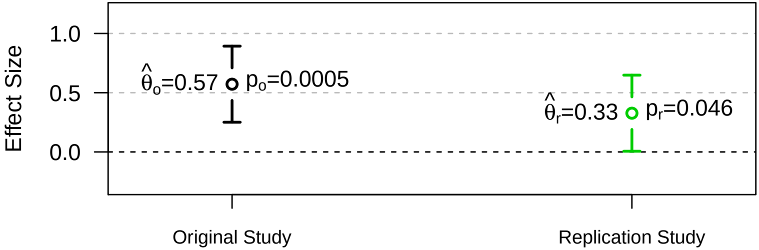
Lack of a single accepted definition



1. Assessment of significance (as in the two-trials rule)
2. Comparison of effect sizes

Replication Success

Lack of a single accepted definition



1. Assessment of significance (as in the two-trials rule)
2. Comparison of effect sizes
3. Meta-analysis combining original and replication effects

A New Standard for the Analysis and Design of Replication Studies

A new standard for the analysis and design of replication studies

Leonhard Held,

University of Zurich, Switzerland

[*Read before* The Royal Statistical Society *at a meeting on 'Signs and sizes: understanding and replicating statistical findings' at the Society's 2019 annual conference in Belfast on Wednesday, September 4th, 2019, the President*, Professor D. Ashby, *in the Chair*]

www.rss.org.uk/Images/PDF/A-new-standard.pdf

A New Standard for the Analysis and Design of Replication Studies

A combination of

- Analysis of Credibility (Matthews, 2001, 2018)
- Assessment of Prior-Data Conflict (Box, 1980)

leads to

1. A new definition of replication success

A New Standard for the Analysis and Design of Replication Studies

A combination of

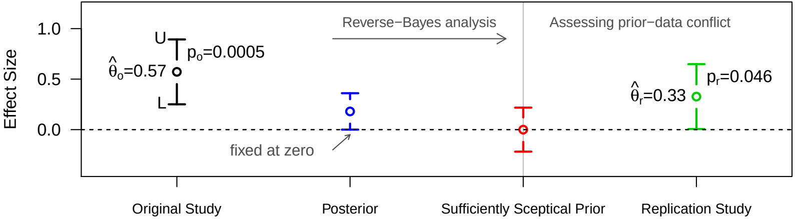
- Analysis of Credibility (Matthews, 2001, 2018)
- Assessment of Prior-Data Conflict (Box, 1980)

leads to

1. A new definition of replication success
2. The sceptical *p*-value to quantify the degree of replication success

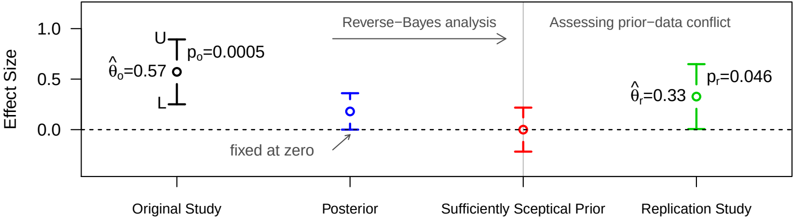
New Definition of Replication Success

1. A sceptic argues, that the original effect $\hat{\theta}_o$, combined with the sufficiently sceptical prior, would no longer be 'significant'.

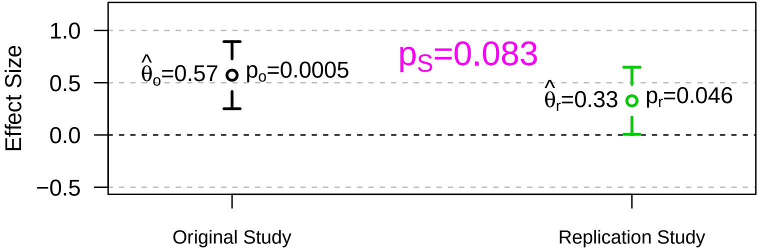


New Definition of Replication Success

1. A sceptic argues, that the original effect $\hat{\theta}_o$, combined with the sufficiently sceptical prior, would no longer be 'significant'.
2. Replication success is declared if the replication effect $\hat{\theta}_r$ is in conflict with the sufficiently sceptical prior.



The Sceptical *P*-Value



If $p_S \leq \alpha$ we have replication success at level α

The Sceptical p-Value

The sceptical p-value $p_S = 2[1 - \Phi(z_S)]$ can be computed from

$$\left(z_o^2/z_S^2 - 1\right)\left(z_r^2/z_S^2 - 1\right) = c,$$

a quadratic equation in z_S^2.

The Sceptical *p*-Value

The sceptical *p*-value $p_S = 2[1 - \Phi(z_S)]$ can be computed from

$$\left(z_o^2/z_S^2 - 1\right)\left(z_r^2/z_S^2 - 1\right) = c,$$

a quadratic equation in z_S^2.

The sceptical *p*-value thus depends on:

$z_o = \hat{\theta}_o/\sigma_o$: Test statistic from original study

The Sceptical *p*-Value

The sceptical *p*-value $p_S = 2[1 - \Phi(z_S)]$ can be computed from

$$\left(z_o^2/z_S^2 - 1\right)\left(z_r^2/z_S^2 - 1\right) = c,$$

a quadratic equation in z_S^2.

The sceptical *p*-value thus depends on:

$z_o = \hat{\theta}_o/\sigma_o$: Test statistic from original study

$z_r = \hat{\theta}_r/\sigma_r$: Test statistic from replication study

The Sceptical *p*-Value

The sceptical *p*-value $p_S = 2[1 - \Phi(z_S)]$ can be computed from

$$\left(z_o^2/z_S^2 - 1\right)\left(z_r^2/z_S^2 - 1\right) = c,$$

a quadratic equation in z_S^2.

The sceptical *p*-value thus depends on:

$z_o = \hat{\theta}_o/\sigma_o$:	Test statistic from original study
$z_r = \hat{\theta}_r/\sigma_r$:	Test statistic from replication study
$c = n_r/n_o$:	Relative sample size

The Sceptical *p*-Value

The sceptical *p*-value $p_S = 2[1 - \Phi(z_S)]$ can be computed from

$$\left(z_o^2/z_S^2 - 1\right)\left(z_r^2/z_S^2 - 1\right) = c,$$

a quadratic equation in z_S^2.

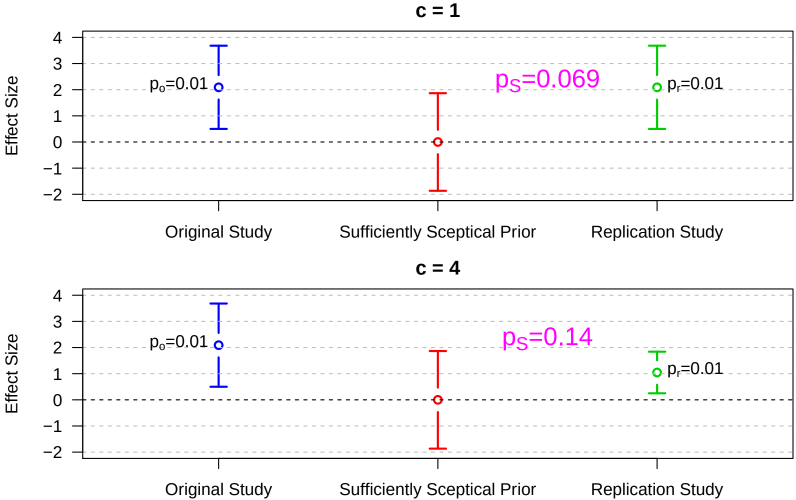
The sceptical *p*-value thus depends on:

$$\begin{aligned} z_o = \hat{\theta}_o/\sigma_o &: \quad \text{Test statistic from original study} \\ z_r = \hat{\theta}_r/\sigma_r &: \quad \text{Test statistic from replication study} \\ c = n_r/n_o &: \quad \text{Relative sample size} \end{aligned}$$

We always have $p_S \geq \max\{p_o, p_r\}$.

Dependence on Relative Sample Size

Both studies significant with $p_o = p_r = 0.01$



Distribution Under the Null

- For $c = 1$, the two studies are treated as exchangeable with $z_S^2 = z_H^2/2$ where z_H^2 is the harmonic mean of the squared z-statistics:

$$z_S^2 = \frac{1}{1/z_o^2 + 1/z_r^2}$$

Distribution Under the Null

- For $c = 1$, the two studies are treated as exchangeable with $z_S^2 = z_H^2/2$ where z_H^2 is the harmonic mean of the squared z-statistics:

$$z_S^2 = \frac{1}{1/z_o^2 + 1/z_r^2}$$

- The null distribution of z_S^2 can be derived.

Distribution Under the Null

– For $c = 1$, the two studies are treated as exchangeable with $z_S^2 = z_H^2/2$ where z_H^2 is the harmonic mean of the squared z-statistics:

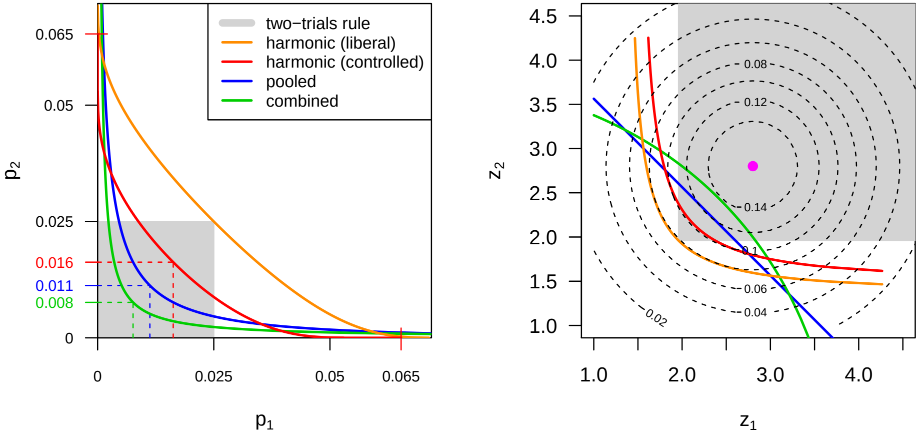
$$z_S^2 = \frac{1}{1/z_o^2 + 1/z_r^2}$$

– The null distribution of z_S^2 can be derived.

→ We can calculate a p-value and a critical value for Type-I error rate control.

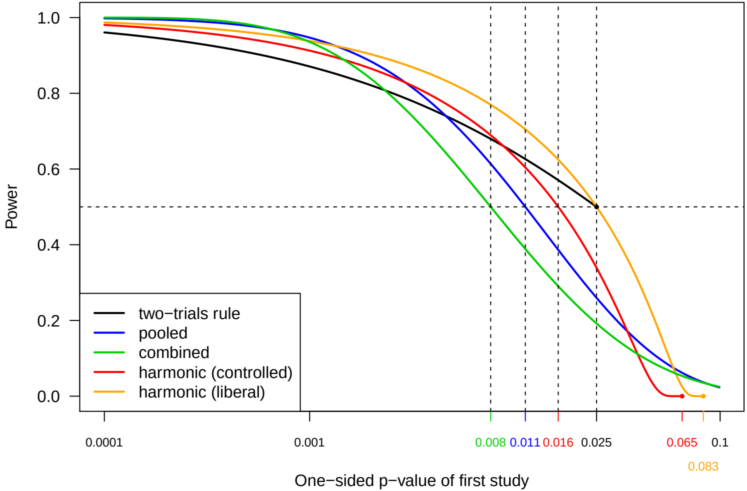
Comparison With the Two-Trials Rule

Type-I error rate control at 0.025^2 except for liberal version



Conditional Power

Power to detect the observed effect from the first study with an identical second study



Project Power

Project power (in %) as a function of the power of the two studies:

Power	two-trials rule	harmonic	combined	pooled
70	49	56	58	61
80	64	71	74	77
90	81	87	90	91
95	90	94	96	97

The Harmonic Mean χ^2 Test

- The approach can be generalized to *n* studies and can also include weights:

$$\chi^2 = \frac{n^2}{\sum_{i=1}^{n} 1/z_i^2} = \frac{n}{z_H^2} \text{ resp. } \chi_w^2 = \frac{w^2}{\sum_{i=1}^{n} w_i/z_i^2} \text{ where } w = \sum_{i=1}^{n} \sqrt{w_i}.$$

The Harmonic Mean χ^2 Test

- The approach can be generalized to n studies and can also include weights:

$$\chi^2 = \frac{n^2}{\sum_{i=1}^{n} 1/z_i^2} = \frac{n}{z_H^2} \text{ resp. } \chi_w^2 = \frac{w^2}{\sum_{i=1}^{n} w_i/z_i^2} \text{ where } w = \sum_{i=1}^{n} \sqrt{w_i}.$$

- The null distribution of χ^2 resp. χ_w^2 can be derived.

The Harmonic Mean χ^2 Test

- The approach can be generalized to n studies and can also include weights:

$$\chi^2 = \frac{n^2}{\sum_{i=1}^{n} 1/z_i^2} = \frac{n}{z_H^2} \text{ resp. } \chi_w^2 = \frac{w^2}{\sum_{i=1}^{n} w_i/z_i^2} \text{ where } w = \sum_{i=1}^{n} \sqrt{w_i}.$$

- The null distribution of χ^2 resp. χ_w^2 can be derived.
- Property of harmonic mean: $z_H^2 \leq n\, z_i^2$ implies bounds on study-specific p-values.

Necessary and Sufficient Bounds

On study-specific p-values at level α_H and n studies

Formalizing the meaning of

> *"at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness"*

α_H	bound	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
1/1600 (two-trials rule)	necessary	0.065	0.17	0.26	0.32	0.37
1/3488556 (five sigma rule)						

Necessary and Sufficient Bounds

On study-specific p-values at level α_H and n studies

Formalizing the meaning of

> *"at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness"*

α_H	bound	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
1/1600 (two-trials rule)	necessary	0.065	0.17	0.26	0.32	0.37
1/3488556 (five sigma rule)	necessary	0.0075	0.058	0.13	0.19	0.24

Necessary and Sufficient Bounds

On study-specific p-values at level α_H and n studies

Formalizing the meaning of

> *"at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness"*

α_H	bound	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
1/1600	necessary	0.065	0.17	0.26	0.32	0.37
(two-trials rule)	sufficient	0.016	0.053	0.099	0.15	0.20
1/3488556	necessary	0.0075	0.058	0.13	0.19	0.24
(five sigma rule)						

Necessary and Sufficient Bounds

On study-specific p-values at level α_H and n studies

Formalizing the meaning of

"at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness"

α_H	bound	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
1/1600	necessary	0.065	0.17	0.26	0.32	0.37
(two-trials rule)	sufficient	0.016	0.053	0.099	0.15	0.20
1/3488556	necessary	0.0075	0.058	0.13	0.19	0.24
(five sigma rule)	sufficient	0.00029	0.0032	0.011	0.024	0.04

Application

Results from 5 clinical trials on the effect of Carvedilol on mortality, from Fisher (1999)

study number	*p*-value	HR	log HR	SE
240	0.0245	0.22	-1.51	0.85
221	0.1305	0.57	-0.56	0.51
220	0.00025	0.27	-1.31	0.41
239	0.2575	0.53	-0.63	1.02
223	0.128	0.72	-0.33	0.29

Application

Results from 5 clinical trials on the effect of Carvedilol on mortality, from Fisher (1999)

study number	*p*-value	HR	log HR	SE
240	0.0245	0.22	-1.51	0.85
221	0.1305	0.57	-0.56	0.51
220	0.00025	0.27	-1.31	0.41
239	0.2575	0.53	-0.63	1.02
223	0.128	0.72	-0.33	0.29

combined $p = 0.00013$

Application

Results from 5 clinical trials on the effect of Carvedilol on mortality, from Fisher (1999)

study number	p-value	HR	log HR	SE
240	0.0245	0.22	-1.51	0.85
221	0.1305	0.57	-0.56	0.51
220	0.00025	0.27	-1.31	0.41
239	0.2575	0.53	-0.63	1.02
223	0.128	0.72	-0.33	0.29

combined $p = 0.00013$

pooled $p = 0.00009$

Application

Results from 5 clinical trials on the effect of Carvedilol on mortality, from Fisher (1999)

study number	*p*-value	HR	log HR	SE
240	0.0245	0.22	-1.51	0.85
221	0.1305	0.57	-0.56	0.51
220	0.00025	0.27	-1.31	0.41
239	0.2575	0.53	-0.63	1.02
223	0.128	0.72	-0.33	0.29

combined $p = 0.00013$

pooled $p = 0.00009$

harmonic $p = 0.00048$

Application

Results from 5 clinical trials on the effect of Carvedilol on mortality, from Fisher (1999)

study number	*p*-value	HR	log HR	SE
240	0.0245	0.22	-1.51	0.85
221	0.1305	0.57	-0.56	0.51
220	0.00025	0.27	-1.31	0.41
239	0.2575	0.53	-0.63	1.02
223	0.128	0.72	-0.33	0.29

combined	$p = 0.00013$
pooled	$p = 0.00009$
harmonic	$p = 0.00048$
weighted harmonic	$p = 0.00034$

Application

Modified data

study number	*p*-value	HR	log HR	SE
240	0.0245	0.22	-1.51	0.85
221	0.1305	0.57	-0.56	0.51
220	0.00025	0.27	-1.31	0.41
239	0.2575	0.53	-0.63	1.02
223	0.256	0.83	-0.19	0.29

Application

Modified data

study number	*p*-value	HR	log HR	SE
240	0.0245	0.22	-1.51	0.85
221	0.1305	0.57	-0.56	0.51
220	0.00025	0.27	-1.31	0.41
239	0.2575	0.53	-0.63	1.02
223	0.256	0.83	-0.19	0.29

combined $p = 0.00021$

Application

Modified data

study number	*p*-value	HR	log HR	SE
240	0.0245	0.22	-1.51	0.85
221	0.1305	0.57	-0.56	0.51
220	0.00025	0.27	-1.31	0.41
239	0.2575	0.53	-0.63	1.02
223	0.256	0.83	-0.19	0.29

combined $p = 0.00021$

pooled $p = 0.00022$

Application

Modified data

study number	*p*-value	HR	log HR	SE
240	0.0245	0.22	-1.51	0.85
221	0.1305	0.57	-0.56	0.51
220	0.00025	0.27	-1.31	0.41
239	0.2575	0.53	-0.63	1.02
223	0.256	0.83	-0.19	0.29

combined $p = 0.00021$

pooled $p = 0.00022$

harmonic $p = 0.0012$

Application

Modified data

study number	*p*-value	HR	log HR	SE
240	0.0245	0.22	-1.51	0.85
221	0.1305	0.57	-0.56	0.51
220	0.00025	0.27	-1.31	0.41
239	0.2575	0.53	-0.63	1.02
223	0.256	0.83	-0.19	0.29

combined	$p = 0.00021$
pooled	$p = 0.00022$
harmonic	$p = 0.0012$
weighted harmonic	$p = 0.0027$

Discussion

“p-values are just too familiar and useful to ditch”

David Spiegelhalter (2017)

Discussion

"p-values are just too familiar and useful to ditch"

David Spiegelhalter (2017)

- The harmonic mean χ^2 test
  - implies restrictions on study-specific *p*-values, requesting each trial to be convincing on its own,

Discussion

"p-values are just too familiar and useful to ditch"
David Spiegelhalter (2017)

- The harmonic mean χ^2 test
  - implies restrictions on study-specific *p*-values, requesting each trial to be convincing on its own,
  - has more power than the two-trials rule,

Discussion

“p-values are just too familiar and useful to ditch”

David Spiegelhalter (2017)

- The harmonic mean χ^2 test
  - implies restrictions on study-specific *p*-values, requesting each trial to be convincing on its own,
  - has more power than the two-trials rule,
  - avoids evidence paradoxes close to the 0.025 threshold,

Discussion

“p-values are just too familiar and useful to ditch”
David Spiegelhalter (2017)

- The harmonic mean χ^2 test
  - implies restrictions on study-specific *p*-values, requesting each trial to be convincing on its own,
  - has more power than the two-trials rule,
  - avoids evidence paradoxes close to the 0.025 threshold,
  - provides a principled extension to analyse results from more than two trials,

Discussion

“p-values are just too familiar and useful to ditch”
David Spiegelhalter (2017)

- The harmonic mean χ^2 test
  - implies restrictions on study-specific *p*-values, requesting each trial to be convincing on its own,
  - has more power than the two-trials rule,
  - avoids evidence paradoxes close to the 0.025 threshold,
  - provides a principled extension to analyse results from more than two trials,
  - and allows for weights.

Discussion

“p-values are just too familiar and useful to ditch”
David Spiegelhalter (2017)

- The harmonic mean χ^2 test
  - implies restrictions on study-specific *p*-values, requesting each trial to be convincing on its own,
  - has more power than the two-trials rule,
  - avoids evidence paradoxes close to the 0.025 threshold,
  - provides a principled extension to analyse results from more than two trials,
  - and allows for weights.
- The sceptical *p*-value

Discussion

"p-values are just too familiar and useful to ditch"
David Spiegelhalter (2017)

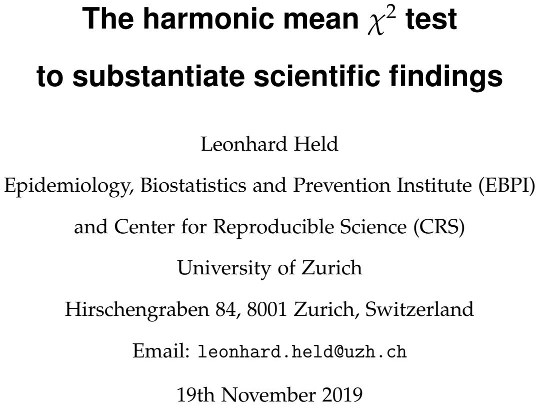
- The harmonic mean χ^2 test
  - implies restrictions on study-specific *p*-values, requesting each trial to be convincing on its own,
  - has more power than the two-trials rule,
  - avoids evidence paradoxes close to the 0.025 threshold,
  - provides a principled extension to analyse results from more than two trials,
  - and allows for weights.
- The sceptical *p*-value
  - can be calibrated to control Type-I error,

Discussion

"p-values are just too familiar and useful to ditch"
David Spiegelhalter (2017)

- The harmonic mean χ^2 test
  - implies restrictions on study-specific *p*-values, requesting each trial to be convincing on its own,
  - has more power than the two-trials rule,
  - avoids evidence paradoxes close to the 0.025 threshold,
  - provides a principled extension to analyse results from more than two trials,
  - and allows for weights.
- The sceptical *p*-value
  - can be calibrated to control Type-I error,
  - may be useful for post-conditional approval studies in "adaptive pathways" for areas of high medical need.

Preprint soon available



The harmonic mean χ^2 test to substantiate scientific findings

Leonhard Held

Epidemiology, Biostatistics and Prevention Institute (EBPI)

and Center for Reproducible Science (CRS)

University of Zurich

Hirschengraben 84, 8001 Zurich, Switzerland

Email: leonhard.held@uzh.ch

19th November 2019