Quantifying Replicability of Multiple Studies in a Meta-Analysis

Mengli Xiao
xiaox345@umn.edu
ENAR 2022

Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455

March 29, 2022

Systematic reviews and meta-analysis: a lens through which evidence is viewed

Image from Murad et al. (2016).

Data integration with meta-analysis

Systematic reviews and meta-analysis

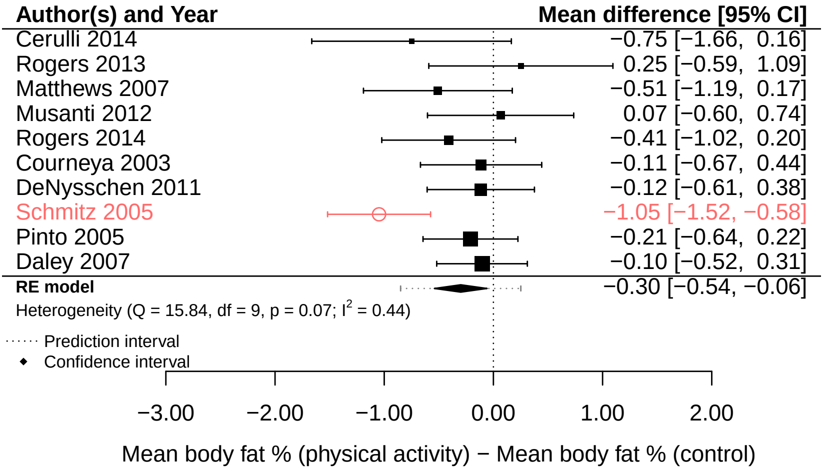
- Totality of evidence.
  – Intervention effects from multiple *related* but *independent* studies $\rightarrow$ one summary effect.
- Crucial tool in evidence-based medicine to justify healthcare decisions.

Why meta-analysis?

- Improved statistical power & precision.
- Investigate reasons for <u>conflicting</u> intervention effects.

Motivating example

- Individual-based activity (left) vs. no exercise (right) and body fat percentage among breast-cancer survivors.



Are studies replicable?

Definition of replicability

"obtaining ***consistent*** results across studies aimed at answering the same scientific question, each of which has obtained its own data." (National Academy of Sciences, 2019)

Why should we assess replicability in meta-analyses?

- Replication crisis of various scientific studies.
- Subjective inclusion and exclusion of studies with different designs.
- Whether a meta-analysis should exclude influential studies?
- Practical guidelines exist, but no consensus on statistical tools.

Current methods for assessing replicability

- Few metrics for replicability in meta-analysis.
- No established definition of replicability.
  - The p-value driven definition is problematic (Jaljuli et al., 2019).

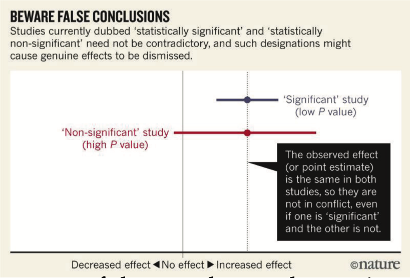


Figure: Using agreement of the p-value to determine replicability in National Academy of Sciences (2019)

Current methods for assessing replicability

- No established definition of replicability.
  - Concept of non-replicability is entangled with heterogeneity (Schauer and Hedges, 2020).
  - Common statistic: τ^2, Q and I^2.
    - τ is the between-study standard deviation.
    - Q and I^2 quantify the heterogeneity ($P(Q) < 0.05$ and $I^2 > 0.75$). $I^2 = \frac{\tau^2}{V_{\text{total}}}$.
  - Question: are Q or I^2 sufficient to reflect non-replicability?

Why are heterogeneity statistics insufficient?

- **High between-study heterogeneity** still assume studies come from the same distribution, i.e., studies are ***replicable***.

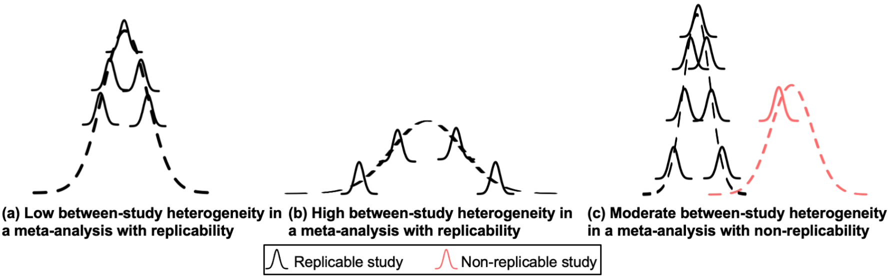


Figure: Random-effects meta-analyses with different levels of heterogeneity and replicability.

Comparison between current methods for assessing replicability

Properties \ Methods	CCMA	Prediction intervals	r-value (CE)	Sceptical p-value	P_{orig}	NCP in Q-statistic	Externalized residual	r-value (RE)	R_m
Symmetric judgement									
Examine whether the first study replicates the following studies	✓	×	✓	×	×	✓	×	✓	✓
Identify non-replicable studies	×	×	×	×	×	×	×	×	✓
Assess effects replicability in an MA	✓	×	✓	×	×	✓	×	✓	✓
Contribution of non-significant study effects									
Use effect sizes (despite statistical significance)	✓	✓	×	✓	✓	✓	✓	×	✓
Allow the overall replicable effect to be null	×	✓	×	×	✓	✓	✓	×	✓
Quantify replicability	×	✓	×	✓	✓	✓	✓	×	✓
Distinguish replicability from homogeneity	×	✓	✓	✓	✓	×	✓	✓	✓
Hypothesis testing									
Null hypothesis is studies replicate	N/A	✓	×	✓	✓	✓	✓	×	✓
First author (year)	Braver (2014)	Patil (2016)	Jaljuli (2021) or Wang (2019)	Held (2020)	Mathur (2020)	Schauer (2020)	Schauer (2020)	Jaljuli (2021)	This paper (202+)

CCMA: Continuously cumulating meta-analysis; CE: Common-effect model; MA: Meta-analysis; N/A: Not applicable; NCP: Non-centrality parameter; RE: Random-effects model.

Comparison between current methods for assessing replicability

Properties \ Methods	CCMA	Prediction intervals	r-value (CE)	Sceptical p-value	P_{orig}	NCP in Q-statistic	Externalized residual	r-value (RE)	R_m
Symmetric judgement									
Examine whether the first study replicates the following studies	✓	×	✓	×	×	✓	×	✓	✓
Identify non-replicable studies	×	×	×	×	×	×	×	×	✓
Assess effects replicability in an MA	✓	×	✓	×	×	✓	×	✓	✓
Contribution of non-significant study effects									
Use effect sizes (despite statistical significance)	✓	✓	×	✓	✓	✓	✓	×	✓
Allow the overall replicable effect to be null	×	✓	×	×	✓	✓	✓	×	✓
Quantify replicability	×	✓	×	✓	✓	✓	✓	×	✓
Distinguish replicability from homogeneity	×	✓	✓	✓	✓	×	✓	✓	✓
Hypothesis testing									
Null hypothesis is studies replicate	N/A	✓	×	✓	✓	✓	✓	×	✓
First author (year)	Braver (2014)	Patil (2016)	Jaljuli (2021) or Wang (2019)	Held (2020)	Mathur (2020)	Schauer (2020)	Schauer (2020)	Jaljuli (2021)	This paper (202+)

CCMA: Continuously cumulating meta-analysis; CE: Common-effect model; MA: Meta-analysis; N/A: Not applicable; NCP: Non-centrality parameter; RE: Random-effects model.

Statistics behind meta-analysis

In a meta-analysis with n studies, if μ is the overall/summary mean effect size, then for each study i (where $i = 1, \ldots, n$):

- y_i: observed effect size; s_i^2: within-study sample variance.
- Goal: obtain an estimate of μ from observed study effects, i.e., $\hat{\mu} = T(y_1, \ldots, y_n, s_1^2, \ldots, s_n^2)$.
- A random-effects model for the observed effect size in each study i:

$$y_i = \mu + \epsilon_i, \qquad \epsilon_i \sim N(0, s_i^2 + \tau^2),$$

where τ^2 is the between-study variance and is estimated from the data.

Statistics behind replicability

Definition of replicability

In order for the summary estimate $\hat{\mu}$ to be a ***consistent*** estimator, studies in a systematic review needs to be ***replicable***.

Meta-analysis summary estimate: $\hat{\mu} = T(y_1, \ldots, y_n, s_1^2, \ldots, s_n^2)$, and assumes $\hat{\mu} \xrightarrow{P} \mu$ as $n \to \infty$

- In the typical random-effects model, the study effect sizes y_i's are:

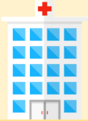
$$y_i = \mu + \epsilon_i, \qquad \epsilon_i \sim N(0, s_i^2 + \tau^2),$$

Intuition

Study 1

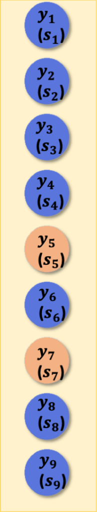
Study 2

⋮



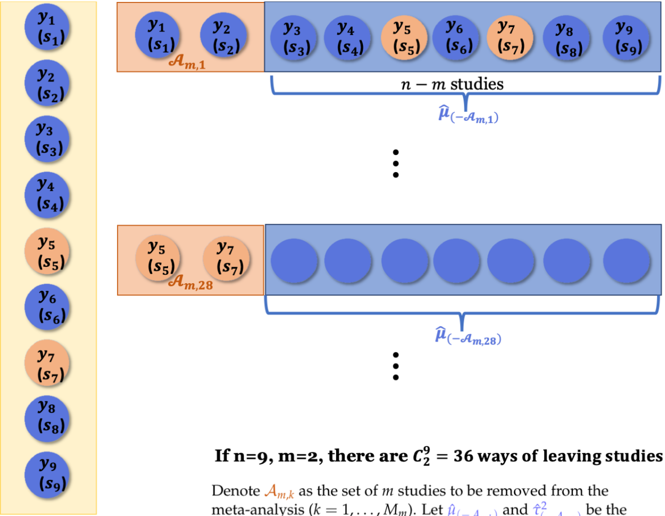
Study 9

Intuition



$\boldsymbol{n}$ studies,
$\boldsymbol{m}$ being non-replicable.

Intuition

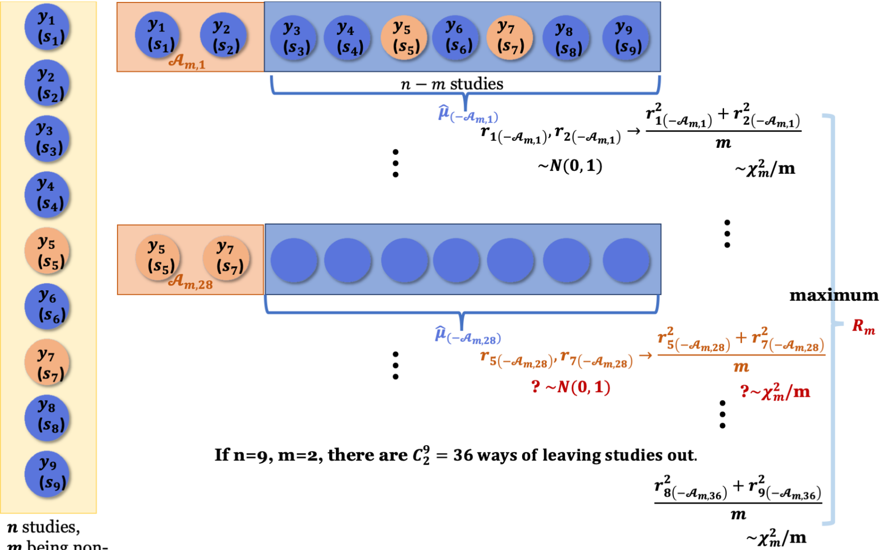


n studies,
m being non-replicable.

If n=9, m=2, there are $C_2^9 = 36$ ways of leaving studies out.

Denote $\mathcal{A}_{m,k}$ as the set of m studies to be removed from the meta-analysis ($k = 1, \ldots, M_m$). Let $\hat{\mu}_{(-\mathcal{A}_{m,k})}$ and $\hat{\tau}^2_{(-\mathcal{A}_{m,k})}$ be the estimates of μ and τ^2 using the **remaining $n-m$ studies**, after omitting studies in $\mathcal{A}_{m,k}$.

Intuition



n studies,
m being non-replicable.

Proposed measure: R_m

- The proposed measure:

$$\boxed{R_m = \max_{k=1,\dots,M_m} R_{\mathcal{A}_{m,k}}}$$

- R_m measures the impact of ***non-replicable*** studies on the total summary effect size.

Properties of R_m

- R_m (maximum of M_m ***dependent*** χ^2_m).
- Proposition 1 (in Appendix) states that R_m (maximum of M_m ***dependent*** χ^2_m) $\xrightarrow{d}$ maximum of M_m ***independent*** χ^2_m
- This applies to y_i's that are not normally distributed.

Asymptotic distribution of R_m – Gumbel distribution

- Under H_0: $y_i \sim N(\mu, s_i^2 + \tau^2)$ for all $i = 1, \ldots, n$, i.e., all studies in the meta-analysis are replicable.
- Proposition 2 (in Appendix) derives the asymptotic distribution of R_m: a Gumbel distribution.
- Based on Proposition 2, the approximate p-value for the replicability test is:

$$P(R_m) \cong 1 - \exp\left(-e^{-c_n^{-1}(mR_m - d_n)}\right).$$

Replicability test

- **Difficulty:** unknown # of non-replicable studies.
- Solution: we assume most studies are replicable in the meta-analysis, and show that it is sufficient to use $m = 1$, i.e., R_1.
- To identify the non-replicable study: an iterative algorithm (in Appendix).

Type I error rate

	$\tau = 0$ ($\tau^2 = 0$)			$\tau = 0.32$ ($\tau^2 = 0.10$)			$\tau = 0.55$ ($\tau^2 = 0.30$)		
		Type I error rate			Type I error rate			Type I error rate	
n	I^2	Gumbel	Bootstrap	I^2	Gumbel	Bootstrap	I^2	Gumbel	Bootstrap
$s_i \sim U(0.54, 1.41)$:									
5	0.14	0.03	0.00	0.17	0.03	0.00	0.23	0.06	0.00
10	0.11	0.03	0.02	0.15	0.04	0.02	0.24	0.08	0.03
20	0.08	0.03	0.03	0.15	0.06	0.05	0.26	0.08	0.06
50	0.06	0.03	0.04	0.13	0.05	0.04	0.27	0.06	0.05
$s_i \sim U(0.22, 0.54)$:									
5	0.14	0.03	0.00	0.34	0.10	0.00	0.56	0.22	0.02
10	0.11	0.03	0.01	0.36	0.12	0.04	0.63	0.19	0.06
20	0.09	0.03	0.02	0.42	0.12	0.07	0.68	0.13	0.06
50	0.06	0.03	0.04	0.44	0.07	0.05	0.71	0.08	0.06
$s_i \sim U(0.10, 0.22)$:									
5	0.14	0.03	0.00	0.70	0.28	0.05	0.87	0.35	0.06
10	0.11	0.03	0.02	0.76	0.20	0.06	0.91	0.20	0.05
20	0.09	0.03	0.02	0.80	0.13	0.05	0.92	0.12	0.05
50	0.06	0.03	0.04	0.82	0.08	0.07	0.93	0.09	0.07

Note: Each setting used 1000 simulated datasets.

Type I error rate

	$\tau = 0$ ($\tau^2 = 0$)			$\tau = 0.32$ ($\tau^2 = 0.10$)			$\tau = 0.55$ ($\tau^2 = 0.30$)		
		Type I error rate			Type I error rate			Type I error rate	
n	I^2	Gumbel	Bootstrap	I^2	Gumbel	Bootstrap	I^2	Gumbel	Bootstrap
$s_i \sim U(0.54, 1.41)$:									
5	0.14	0.03	0.00	0.17	0.03	0.00	0.23	0.06	0.00
10	0.11	0.03	0.02	0.15	0.04	0.02	0.24	0.08	0.03
20	0.08	0.03	0.03	0.15	0.06	0.05	0.26	0.08	0.06
50	0.06	0.03	0.04	0.13	0.05	0.04	0.27	0.06	0.05
$s_i \sim U(0.22, 0.54)$:									
5	0.14	0.03	0.00	0.34	0.10	0.00	0.56	0.22	0.02
10	0.11	0.03	0.01	0.36	0.12	0.04	0.63	0.19	0.06
20	0.09	0.03	0.02	0.42	0.12	0.07	0.68	0.13	0.06
50	0.06	0.03	0.04	0.44	0.07	0.05	0.71	0.08	0.06
$s_i \sim U(0.10, 0.22)$:									
5	0.14	0.03	0.00	0.70	0.28	0.05	0.87	0.35	0.06
10	0.11	0.03	0.02	0.76	0.20	0.06	0.91	0.20	0.05
20	0.09	0.03	0.02	0.80	0.13	0.05	0.92	0.12	0.05
50	0.06	0.03	0.04	0.82	0.08	0.07	0.93	0.09	0.07

Note: Each setting used 1000 simulated datasets.

Type I error rate

	$\tau = 0$ ($\tau^2 = 0$)			$\tau = 0.32$ ($\tau^2 = 0.10$)			$\tau = 0.55$ ($\tau^2 = 0.30$)		
		Type I error rate			Type I error rate			Type I error rate	
n	I^2	Gumbel	Bootstrap	I^2	Gumbel	Bootstrap	I^2	Gumbel	Bootstrap
$s_i \sim U(0.54, 1.41)$:									
5	0.14	0.03	0.00	0.17	0.03	0.00	0.23	0.06	0.00
10	0.11	0.03	0.02	0.15	0.04	0.02	0.24	0.08	0.03
20	0.08	0.03	0.03	0.15	0.06	0.05	0.26	0.08	0.06
50	0.06	0.03	0.04	0.13	0.05	0.04	0.27	0.06	0.05
$s_i \sim U(0.22, 0.54)$:									
5	0.14	0.03	0.00	0.34	0.10	0.00	0.56	0.22	0.02
10	0.11	0.03	0.01	0.36	0.12	0.04	0.63	0.19	0.06
20	0.09	0.03	0.02	0.42	0.12	0.07	0.68	0.13	0.06
50	0.06	0.03	0.04	0.44	0.07	0.05	0.71	0.08	0.06
$s_i \sim U(0.10, 0.22)$:									
5	0.14	0.03	0.00	0.70	0.28	0.05	0.87	0.35	0.06
10	0.11	0.03	0.02	0.76	0.20	0.06	0.91	0.20	0.05
20	0.09	0.03	0.02	0.80	0.13	0.05	0.92	0.12	0.05
50	0.06	0.03	0.04	0.82	0.08	0.07	0.93	0.09	0.07

Note: Each setting used 1000 simulated datasets.

Power (detect non-replicability)

Similar power to identify the right non-replicable study

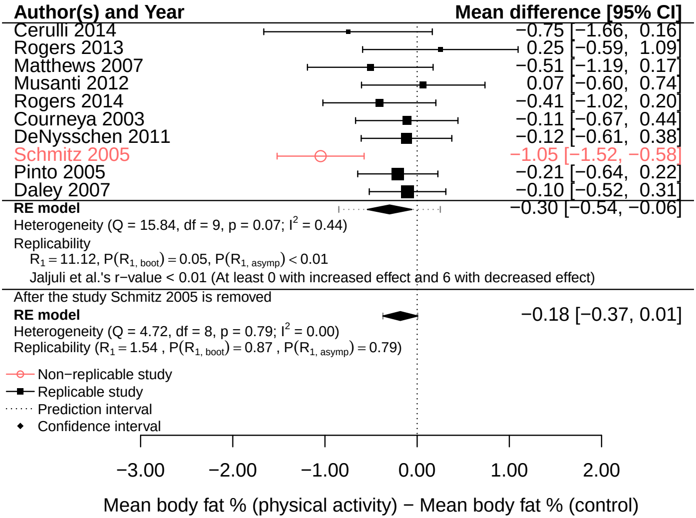
	$\tau = 0$ ($\tau^2 = 0$)			$\tau = 0.32$ ($\tau^2 = 0.10$)			$\tau = 0.55$ ($\tau^2 = 0.30$)		
		Power			Power			Power	
n	I^2_{tot} (I^2_{rep})	Gumbel	Bootstrap	I^2_{tot} (I^2_{rep})	Gumbel	Bootstrap	I^2_{tot} (I^2_{rep})	Gumbel	Bootstrap
$s_i \sim U(0.54, 1.41)$:									
5	0.56 (0.15)	0.45	0.08	0.57 (0.17)	0.43	0.07	0.59 (0.23)	0.42	0.07
10	0.42 (0.11)	0.47	0.32	0.45 (0.15)	0.45	0.28	0.50 (0.23)	0.38	0.22
20	0.29 (0.09)	0.48	0.43	0.36 (0.15)	0.44	0.37	0.44 (0.26)	0.38	0.26
50	0.17 (0.06)	0.44	0.44	0.24 (0.13)	0.40	0.38	0.36 (0.27)	0.35	0.27
$s_i \sim U(0.22, 0.54)$:									
5	0.92 (0.15)	1.00	0.73	0.92 (0.32)	0.96	0.51	0.92 (0.53)	0.84	0.30
10	0.86 (0.11)	1.00	0.99	0.87 (0.36)	0.98	0.92	0.89 (0.61)	0.88	0.66
20	0.75 (0.09)	1.00	1.00	0.80 (0.41)	0.99	0.97	0.85 (0.68)	0.86	0.83
50	0.53 (0.06)	1.00	1.00	0.68 (0.44)	0.98	0.98	0.79 (0.71)	0.85	0.90
$s_i \sim U(0.10, 0.22)$:									
5	0.99 (0.15)	1.00	1.00	0.99 (0.66)	1.00	0.79	0.99 (0.84)	0.93	0.36
10	0.97 (0.11)	1.00	1.00	0.98 (0.75)	1.00	1.00	0.98 (0.90)	0.96	0.88
20	0.95 (0.09)	1.00	1.00	0.96 (0.80)	1.00	1.00	0.97 (0.92)	0.97	0.99
50	0.89 (0.06)	1.00	1.00	0.93 (0.82)	1.00	1.00	0.96 (0.93)	0.96	1.00

Note: Each setting used 1000 simulated datasets.
I^2_{tot} is the observed I^2 statistic for all studies in a meta-analysis.
I^2_{rep} is the observed I^2 statistic for replicable studies only.

Can physical activity reduce body fat among breast-cancer survivors?

- Meta-analysis in Lahart et al. (2018) with 10 studies: individual-format physical activity versus control for reducing body fat in women with breast cancer after adjuvant therapy (chemotherapy and/or radiation therapy).
- Mean difference between groups is the measure, a continuous outcome.

Can physical activity reduce body fat percentage among breast-cancer survivors?



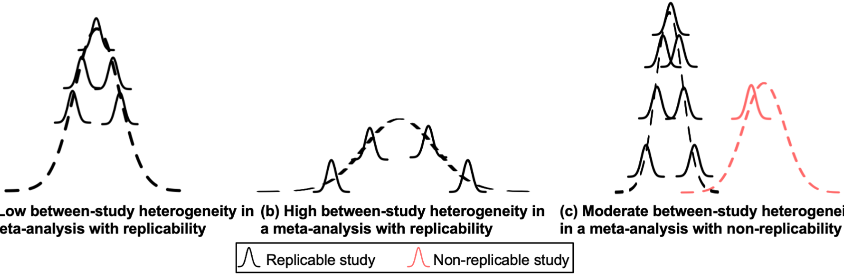
Can physical activity reduce body fat percentage among breast-cancer survivors?

- Schmitz 2005 only included participants who completed a college education and the intervention only used resistance training.
- This may suggest further research questions in how physical activity can reduce body fat percentage to improve patient care.

1. Introduction
   - Motivating example and replicability
   - Methods
     - Meta-analysis model
     - Proposed replicability measure
     - Replicability test
   - Simulation results
     - Type I error rate
     - Power
   - Case studies
   - Discussion

Contributions

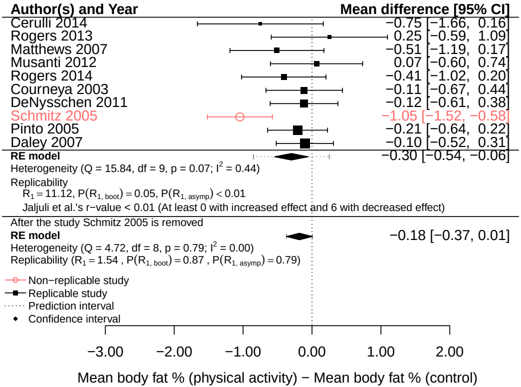
- New metric for meta-analysis.
  - Quantify replicability regardless of heterogeneity.



(a) Low between-study heterogeneity in a meta-analysis with replicability (b) High between-study heterogeneity in a meta-analysis with replicability (c) Moderate between-study heterogeneity in a meta-analysis with non-replicability

Contributions

- New metric for meta-analysis.
  - Apply to non-significant effects.



Contributions

- The new metric can suggest sources of effect discrepancy in a meta-analysis: confounding, selection bias, measurement error, etc.
    - Improve future systematic reviews.
- To facilitate building consensus in any research topic and reduce replication crisis, we developed an R package "repMeta".

Future directions

- Method development for integrating complex dataset.
  - Multiple treatments, e.g., network meta-analysis (Lin et al., 2016).
  - Multiple outcomes, e.g., multivariate meta-analysis, longitudinal outcomes (Riley et al., 2017).
  - Multiple covariates — meta-regression approach.
  - Individual participant data, e.g., hierarchical models for meta-analysis (Abo-Zaid et al., 2013).
- Applications.
  - Evaluate the robustness of overall effect from multi-center clinical trials and observation studies, and find factors that cause inconsistency, e.g., education.

$\vdots$

Thank you!

Questions? (`xiaox345@umn.edu`)

References: I

Abo-Zaid, G., Guo, B., Deeks, J. J., Debray, T. P., Steyerberg, E. W., Moons, K. G., and Riley, R. D. (2013). Individual participant data meta-analyses should not ignore clustering. *Journal of clinical epidemiology*, 66(8):865–873.

Embrechts, P., Mikosch, T., and Klüppelberg, C. (1997). *Modelling Extremal Events: For Insurance and Finance*. Springer-Verlag, Berlin, Germany.

Gumbel, E. J. (1958). *Statistics of Extremes.* Columbia University Press, New York, NY.

Jaljuli, I., Benjamini, Y., Shenhav, L., Panagiotou, O., and Heller, R. (2019). Quantifying replicability and consistency in systematic reviews. *ArXiv.* Available at `https://arxiv.org/abs/1907.06856`.

References: II

Lahart, I. M., Metsios, G. S., Nevill, A. M., and Carmichael, A. R. (2018). Physical activity for women with breast cancer after adjuvant therapy. *Cochrane Database of Systematic Reviews*, 1:Art. No.: CD011292.

Lin, L., Chu, H., and Hodges, J. S. (2016). Sensitivity to excluding treatments in network meta-analysis. *Epidemiology*, 27(4):562–569.

Murad, M. H., Asi, N., Alsawas, M., and Alahdab, F. (2016). New evidence pyramid. *BMJ Evidence-Based Medicine*, 21(4):125–127.

National Academy of Sciences (2019). *Reproducibility and Replicability in Science*. National Academies Press, Washington, DC.

Riley, R. D., Jackson, D., Salanti, G., Burke, D. L., Price, M., Kirkham, J., and White, I. R. (2017). Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. *BMJ*, 358.

References: III

Schauer, J. M. and Hedges, L. V. (2020). Assessing heterogeneity and power in replications of psychological experiments. *Psychological Bulletin*, 146:701–719.

Proposition 1

Proposition 1

In a collection of n studies, for $i = 1, \ldots, n$, assume that all studies with effect sizes y_i's replicate each other by sharing a common overall mean μ, and the fourth moments of y_i's are finite. Then, in the leave-m-studies-out procedure for assessing replicability, $R_m \xrightarrow{d} m^{-1} \max_{k=1,\ldots,M_m} C_k$ as $M_m \to \infty$, where $C_k \overset{\text{iid}}{\sim} \chi^2_m$.

Proposition 2

- Under H_0: $y_i \sim N(\mu, s_i^2 + \tau^2)$ for all $i = 1, \ldots, n$, i.e., all studies in the meta-analysis are replicable

Proposition 2

(Gumbel, 1958; Embrechts et al., 1997)

Under H_0, as $M_m = \mathcal{C}_m^n \to \infty$, $-c_n^{-1}(mR_m - d_n) \xrightarrow{d} G$, where $c_n = 2$, $d_n = 2\left[\log(\mathcal{C}_m^n) + (m/2 - 1)\log\log(\mathcal{C}_m^n) - \log\Gamma(m/2)\right]$, and G is a standard Gumbel random variable with CDF $F_G(x) = \exp(-e^{-x})$.

- Based on Proposition 2, the approximate p-value for the replicability test is:

$$P(R_m) \cong 1 - \exp\left(-e^{-c_n^{-1}(mR_m - d_n)}\right).$$

Algorithm

Algorithm 1: Procedure for identifying non-replicable studies based on R_1.

Result: Indexes of non-replicable studies

Using Equation (2.5), calculate $M_1 = n$ values of $R_{\mathscr{A}_{1,(k)}}$ for $k = 1, \ldots, n$;

Order them such that $R_{\mathscr{A}_{1,(1)}} < R_{\mathscr{A}_{1,(2)}} < \cdots < R_{\mathscr{A}_{1,(n)}}$;

Let $l \leftarrow n$;

while $P\left(\max_{k=1,\ldots,l} R_{\mathscr{A}_{1,(k)}}\right) < \alpha$ **do**

Pick the two largest values of $R_{\mathscr{A}_{1,(k)}}$, i.e., $R_{\mathscr{A}_{1,(l-1)}}$ and $R_{\mathscr{A}_{1,(l)}}$;

Leave the corresponding studies out, indexed by a and b;

Obtain the sets of studies $\{1, \ldots, l\} \setminus a$ and $\{1, \ldots, l\} \setminus b$;

Compare $\max_{k \in \{1, \ldots, l\} \setminus a} R_{\mathscr{A}_{1,(k)}}$ with $\max_{k \in \{1, \ldots, l\} \setminus b} R_{\mathscr{A}_{1,(k)}}$;

Let $m \leftarrow \operatorname{argmin}_{\{a,b\}} \left\{ \max_{k \in \{1, \ldots, l\} \setminus a} R_{\mathscr{A}_{1,(k)}}, \max_{k \in \{1, \ldots, l\} \setminus b} R_{\mathscr{A}_{1,(k)}} \right\}$;

Omit study m and let $l \leftarrow l - 1$;

Using Equation (2.5), calculate $M_1 = l$ values of $R_{\mathscr{A}_{1,(k)}}$ for $k = 1, \ldots, l$;

Order them such that $R_{\mathscr{A}_{1,(1)}} < R_{\mathscr{A}_{1,(2)}} < \cdots < R_{\mathscr{A}_{1,(l)}}$;

end