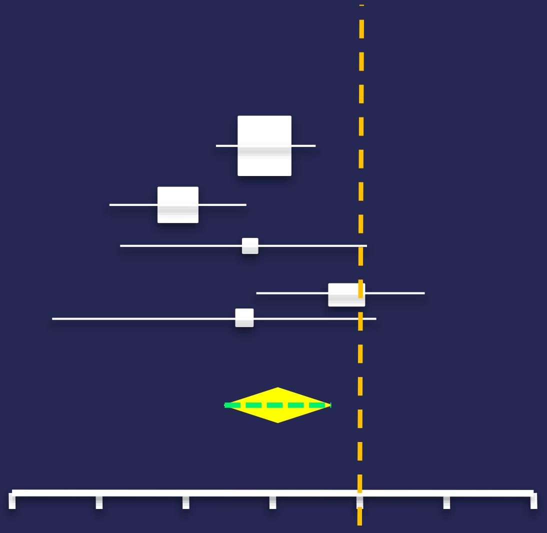
Methods to calculate the uncertainty in the estimated overall effect size under the random-effects model



Areti Angeliki Veroniki, PhD

Prepared for: *Edinburgh Cochrane Colloquium: SMG scientific meeting*

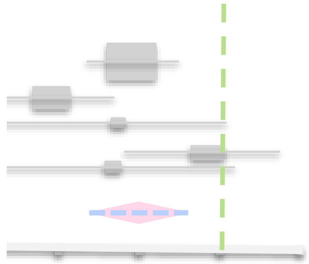
September 17, 2018

School of Education,
University of Ioannina,
Ioannina,
Greece

EVIDENCE SYNTHESIS METHODS STATISTICS TEAM

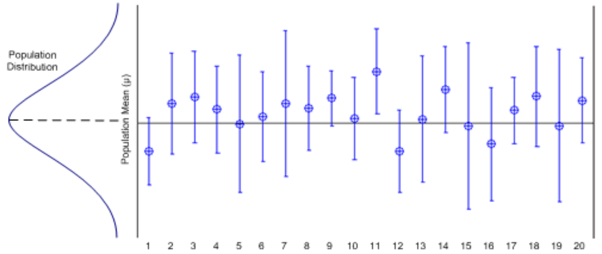
Cochrane

I have no actual or potential conflict of interest in relation to this presentation



Aims of the presentation

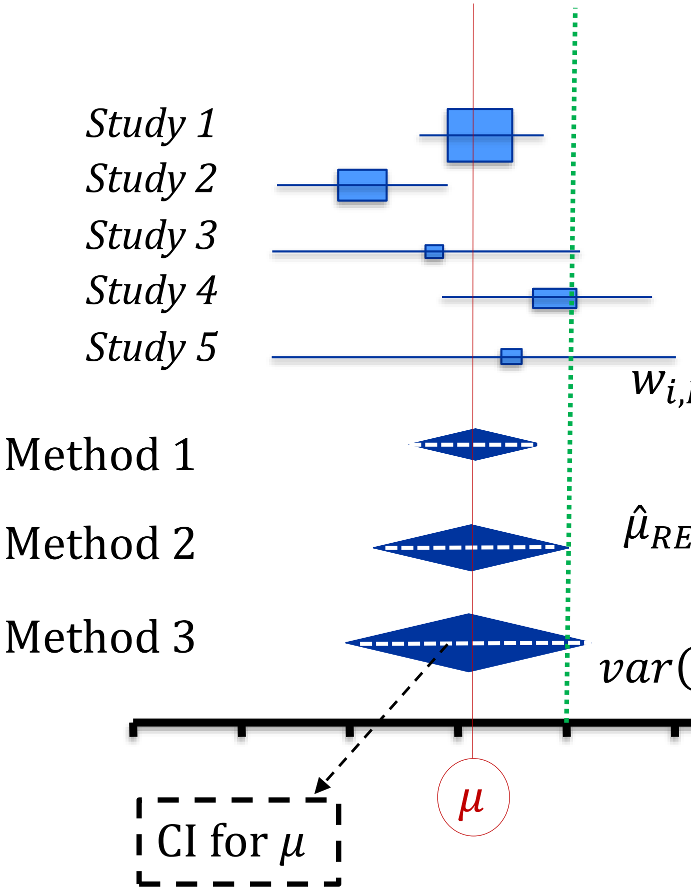
- To summarize different methods to calculate uncertainty in the estimated overall effect size under the random-effects model.
  - Can different methods impact our decision-making?
- To discuss how different methods to calculate the uncertainty in the estimated overall effect size can affect meta-analysis' results.
  - What are the properties of the different methods?
- To present real-life and simulation findings for calculating confidence intervals and prediction intervals for the overall effect size.
  - Which method is the most appropriate to apply?
- To identify factors that may control the calculation of a confidence interval by considering the results of comparative simulation and real-life data studies.
  - Which methods are preferable than others and under which circumstances?



Meta-analysis

Which is the most appropriate method to use?

- Plethora of methods exist to calculate uncertainty in the estimated overall effect size.
- The performance of a method may vary in various meta-analysis settings.
- The choice of the method calculation of uncertainty in the estimated overall effect size is important when conducting a meta-analysis.
- An erroneous choice of the method could lead to misleading results.



$$y_i = \theta_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, v_i)$$

$$\theta_i \sim N(\mu, \tau^2)$$

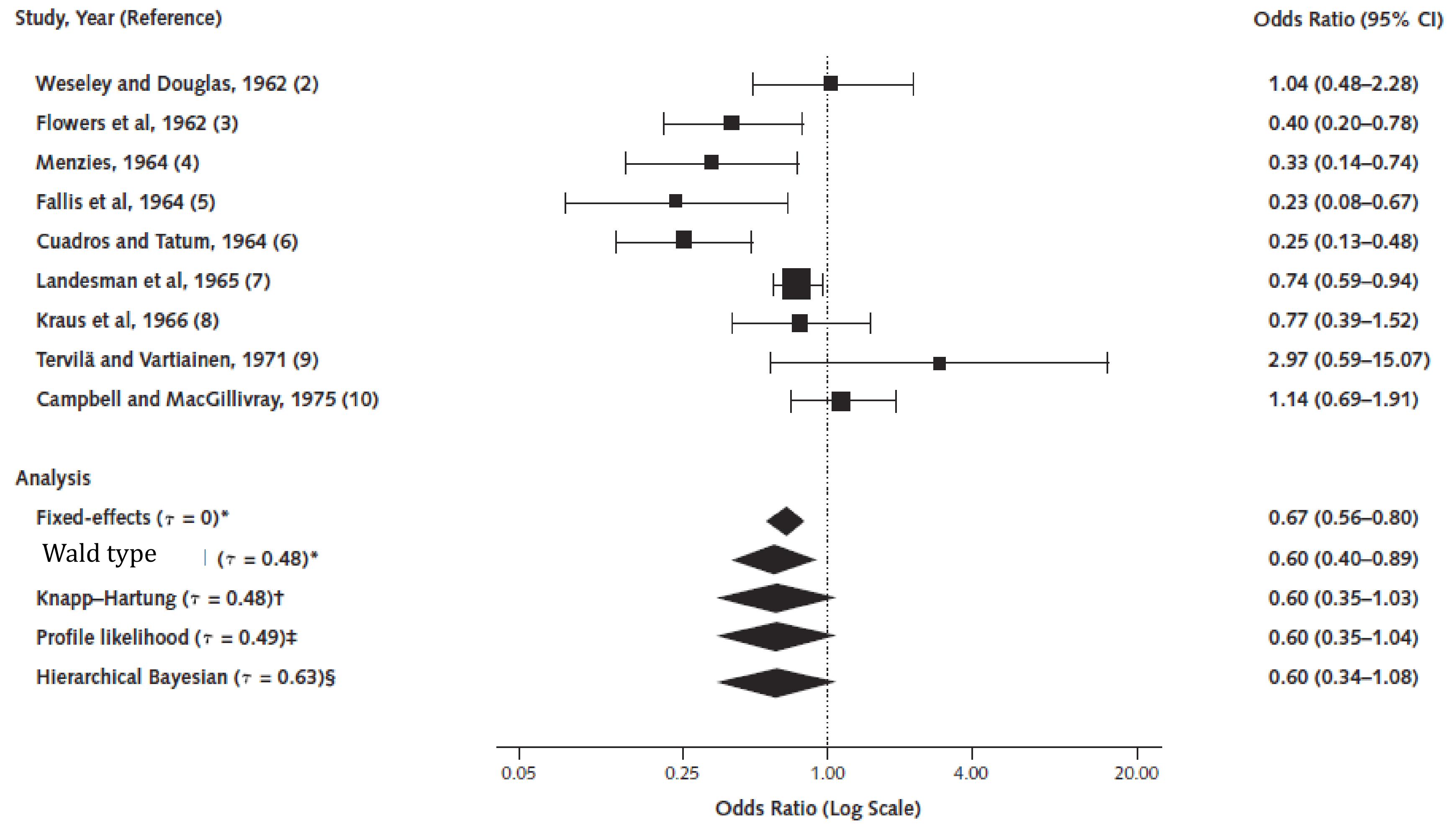
$$w_{i,RE} = \frac{1}{v_i + \hat{\tau}^2}$$

$$\hat{\mu}_{RE} = \frac{\sum y_i w_{i,RE}}{\sum w_{i,RE}}$$

$$var(\hat{\mu}_{RE}) = \frac{1}{\sum w_{i,RE}}$$

Various CIs can lead to different conclusions

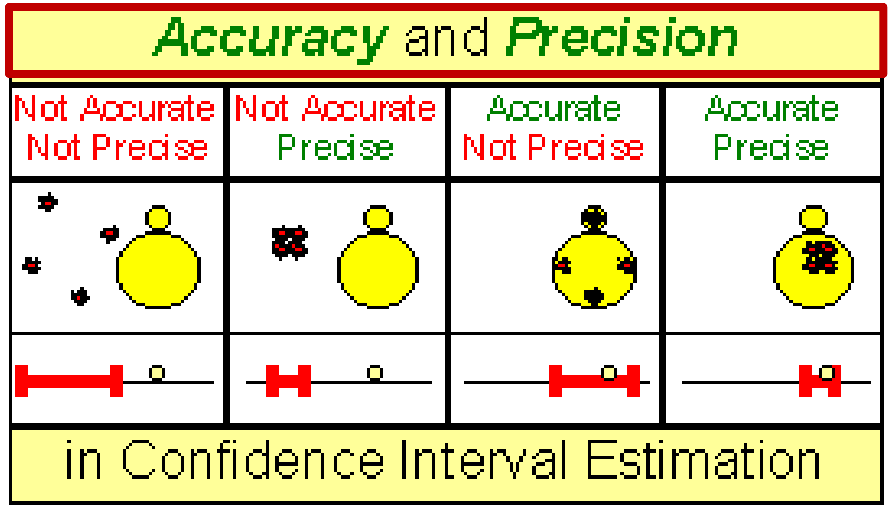
Figure. Heterogeneous evidence from Collins and colleagues' meta-analysis of the effects of diuretics on preeclampsia (11).



Cornel et al. Annals of Internal Medicine 2014

Confidence Interval (CI) for the overall effect size

Estimation of summary effect size
- Point Estimation
- Interval Estimation



Desirable properties

- ✓ Accuracy = High Coverage Probability – P(μ∈ CI)
  - The closer the coverage is to the nominal level (usually 0.95) the better the CI.
- ✓ Precision = Narrow CI
  - Narrower CIs retaining the correct coverage are preferable because they increase precision.

Department of Primary Education, School of Education, University of Ioannina, Ioannina, Greece

Literature Review of CI methods

Our search identified:

- 69 relevant publications
- 15 methods to compute a CI for the overall effect size (grouped in 7 broad categories).

The properties of the methods were evaluated in 31 papers:

- including 30 simulation studies and 32 real-life data evaluations of ≥2methods.

The most popular technique is WTz

Veroniki et al. Res Synth Methods. 2018. doi: 10.1002/jrsm.1319.

Categories

A. Wald-type (WT) CIs
    a) Wald-type normal distribution (WTz)
    b) Wald-type t-distribution (WTt)
    c) Quantile approximation (WTqa)

B. Hartung-Knapp/Sidik-Jonkman (HKSJ) CIs

C. Likelihood-based CIs
    a) Profile likelihood (PL)
    b) Higher-order likelihood inference methods

D. Henmi and Copas (HC) CIs

E. Biggerstaff and Tweedie (BT) CIs

F. Resampling CIs
    a) Zeng and Lin (ZL)
    b) Bootstrap
    c) Follmann and Proschan (FP)

G. Bayesian Credible Intervals

Confidence Interval methods

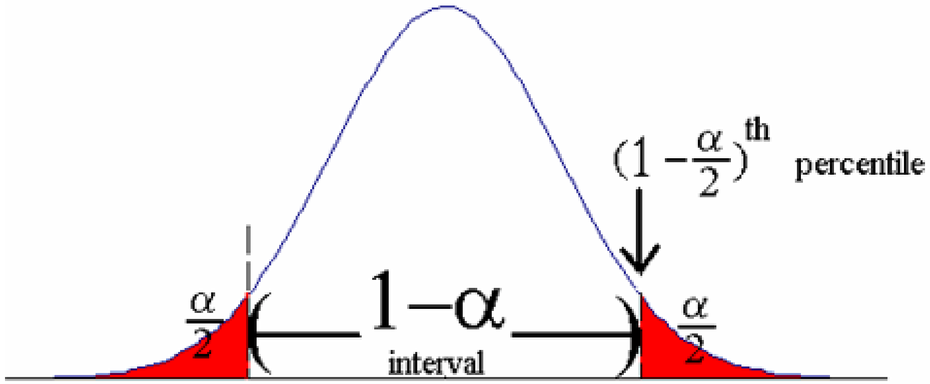
No	Method	Confidence Interval
1	Wald-type normal distribution (WTz)	$\hat{\mu}_{RE} \pm z_{0.975}\sqrt{var(\hat{\mu}_{RE})}$
2	Wald-type t-distribution (WTt)	$\hat{\mu}_{RE} \pm t_{k-1,0.975}\sqrt{var(\hat{\mu}_{RE})}$
3	Quantile approximation (WTqa)	$\hat{\mu}_{RE} \pm b_k\sqrt{var(\hat{\mu}_{RE})}$, with b_k the quantile approximation function of the distribution of the statistic $M = \frac{\hat{\mu}_{RE}-\mu}{\sqrt{var(\hat{\mu}_{RE})}}$
4	Hartung-Knapp/Sidik-Jonkman (HKSJ)	$\hat{\mu}_{RE} \pm t_{k-1,0.975}\sqrt{\sigma^2_{w,\hat{\mu}_{RE}}}$, with $\sigma^2_{w,\hat{\mu}_{RE}} = q \cdot var(\hat{\mu}_{RE})$, $q = \frac{Q_{gen}}{k-1}$, and $Q_{gen} = \sum w_{i,RE}(y_i - \hat{\mu}_{RE})^2$
5	Modified HKSJ	HKSJ, but use q^* instead of q: $q^* = \max\{1, q\}$
6	Profile likelihood (PL)	Profile log-likelihood for μ: $lnL_p(\mu) = lnL(\mu, \hat{\tau}^2_{ML}(\mu))$, $lnL_p(\mu) > lnL_p(\hat{\mu}_{RE}) - \frac{\chi^2_{1,0.05}}{2}$

Confidence Interval methods

No	Method	Confidence Interval
7, 8	Higher-order likelihood inference methods	The Bartlett-type adjusted efficient score statistic (BES) (No 7) and Skovgaard's statistic (SS) (No 8) use a higher-order approximation than the PL
9	Henmi and Copas (HC)	Hybrid approach: the FE estimate is accompanied by a CI that allows for τ^2 under the assumptions of a RE model
10	Biggerstaff and Tweedie (BT)	$\hat{\mu}_{RE}^{BT} \pm z_{0.975}\sqrt{var(\hat{\mu}_{RE}^{BT})}$, with $var(\hat{\mu}_{RE}^{BT}) = \frac{1}{(\sum w_{i,RE}^{BT})^2}\sum\left(w_{i,RE}^{BT}\right)^2(v_i + \hat{\tau}^2)$ and $w_{i,RE}^{BT} = E\left(w_{i,RE}\right)$
11	Resampling methods: Zeng and Lin (ZL)	Simulate values of τ^2 using DL, then simulate estimated average effect sizes using the sampled τ^2 to calculate the weights in $\hat{\mu}_{RE} = \frac{\sum y_i w_{i,RE}}{\sum w_{i,RE}}$. Repeat both aspects B times, get empirical distribution of $\hat{\mu}_{RE}$ and compute CI
12, 13	Resampling methods: Bootstrap confidence intervals	Non-parametric bootstrap CI (No 12) with resampling from the sample itself with replacement, and Parametric bootstrap CI (No 13) with resampling from a fitted model

Confidence Interval methods

No	Method	Confidence Interval
14	Resampling methods: Follmann and Proschan (FP)	Permutation tests can be extended to calculate CIs for the effect size. CIs are constructed by inverting hypothesis test to give the CI bounds - parameter values that are not rejected by the hypothesis test lie within the corresponding CI
15	Bayesian credible intervals	Bayesian credible intervals for the overall effect size can be obtained within a Bayesian framework



Comparative evaluation of the methods

i. Wald-type methods (WTz, WTt, WTqa)

- ☑ For large number of studies WTz, WTt, and WTqa perform well. [1, 2]
- ☒ WTz performs worse in terms of coverage for small number of studies (k<16) compared with the PL and the WTt methods. [1]
- ☒ WTz and WTt depend on the number of studies, the τ^2 estimator, and the τ^2 magnitude. [4]
- ☒ Coverage of WTz has been found to be as low as 65% (at 95% nominal level) when I^2=90% and k=2,3. [3]
- ☒ Coverage of WTt may be below the 95% nominal level, but it becomes conservative (close to 1) when k is small. [1, 2, 3]
- ☒ WTqa and WTt have on average similar coverage, but WTqa outperforms WTz, PL, and ZL CIs – but it is very conservative. [2, 6]
- ☒ WTqa has been criticized that it is very difficult to obtain suitable critical values b_k that apply to all meta-analyses. [5]

Implement in RevMan?	
WTz	Implemented
WTt	☒
WTqa	☒

WTz: Wald type – normal distr

WTt: Wald type – t distr

WTqa: Wald type – quantile approximation

1: Jackson et al J Stat Plan Infer 2010, 2: Brockwell and Gordon Stat Med 2007, 3: Langan et al RSM 2018, 4: Sanchez-Meca and Marin-Martinez Psychol Methods 2008, 5: Jackson and Bowden Stat Med. 2009, 6: Zeng and Lin Biometrika. 2015

Comparative evaluation of the methods

ii. Hartung-Knapp/Sidik-Jonkman methods (HKSJ, modified HKSJ)

- ☑ HKSJ on average produces wider CIs with more coverage than the WTz and WTt methods.[1, 2, 3]
- ☑ HKSJ has coverage close to the nominal level, is not influenced by the magnitude or estimator of τ^2, and is insensitive to the number of trials. [1, 2, 3, 4, 5]
- ☑ Simulations suggest HKSJ has good coverage for the odds ratio, risk ratio, mean difference, and standardized mean difference effect measures. [3, 7]
- ☐ Real-life data studies showed that the WTz method yielded more often statistically significant results compared with the HKSJ method. [1, 6]
- ☒ HKSJ is suboptimal than the WTz and WTt CIs when binary outcomes with rare events are included in a meta-analysis. [2]
- ☒ Caution is needed for the HSKJ CI when <5 studies of unequal sizes are included in a meta-analysis. [4, 6]
- ☒ In the absence of heterogeneity it may be: HKSJ coverage < WTz coverage. [1]

1:IntHout et al BMC Med Res Methodol. 2014, 2: Langan et al RSM 2018, 3: Makambi J Biopharm Stat. 2004, 4: Hartung Biom J 1999, 5: Sanchez-Meca and Marin-Martinez Psychol Methods 2008, 6: Wiksten et al Stat Med. 2016, 7: Sidik and Jonkman Stat Med. 2002

WTz: Wald type – normal distr

WTt: Wald type – t distr

Comparative evaluation of the methods

ii. Hartung-Knapp/Sidik-Jonkman methods (HKSJ, modified HKSJ)

☑ The modified HKSJ is preferable when few studies of varying size and precision are available. [1]

☒ For small k (particularly for k=2) and small τ^2 the modified HKSJ tends to be over-conservative. [1, 2, 3]

Implement in RevMan?	
HKSJ	☑
mHKSJ	☑

1: Röver et al BMC Med Res Methodol. 2015, 2: Jackson et al Stat Med. 2017, 3: Viechtbauer Psychol Methods. 2015, 4: Brockwell and Gordon Stat Med. 2007, 5: Kosmidis Biometrika. 2017, 6: Noma Stat Med 2011, 7: Guolo & Varin Stat Methods Med Res. 2015

Comparative evaluation of the methods

iii. Likelihood-based methods (PL, BES, SS)

- ☑ PL has higher coverage closer to the nominal level than WTz and WTt, even when k is relatively small ($k \leq 8$). [4, 5]
- ☑ BES improves coverage over WTz, WTt, and PL CIs as τ^2 increases and/or k decreases. [6]
- ☑ SS yields similar results with BES, and has better coverage than WTz and PL CIs. [6, 7]
- ☒ Caution is needed for k≤5 as BES tends to be over-conservative. [6]

Implement in RevMan?	
PL	?
BES	?
SS	?

WTz: Wald type – normal distr

WTt: Wald type – t distr

PL: Profile Likelihood

BES: Bartlett-type adjusted efficient score statistic

SS: Skovgaard's statistic

1: Röver et al BMC Med Res Methodol. 2015, 2: Jackson et al Stat Med. 2017, 3: Viechtbauer Psychol Methods. 2015, 4: Brockwell and Gordon Stat Med. 2007, 5: Kosmidis Biometrika. 2017, 6: Noma Stat Med 2011, 7: Guolo & Varin Stat Methods Med Res. 2015

Comparative evaluation of the methods

iv. Henmi and Copas method (HC)

- ☑ For k>10 HC yields better coverage than WTz, HKSJ, PL, and BT methods, irrespective the absence/presence of publication bias . [1]
- ☒ For k<10 the HKSJ and PL methods perform better than HC, WTz, and BT methods.[1]

v. Biggerstaff and Tweedie method (BT)

- ☒ WTz and BT methods have comparable coverage (below the nominal level), but coverage increases for the exact weights. [2, 3]

vi. Resampling methods (ZL, FP)

- ☑ ZL outperforms both WTz and PL for small k in terms of coverage. [4]
- ☑ FP controls coverage better than WTz, WTt, PL, and is closely followed by BES. [5]
- ☒ BES is slightly more powerful than FP especially for small k. [5]

Implement in RevMan?	
HC	☒
BT	☒
ZL	?
FP	?

WTz: Wald type – normal distr
WTt: Wald type – t distr
HKSJ: Hartung-Knapp/Sidik-Jonkman
PL: Profile Likelihood
BES: Bartlett-type adj score statistic
ZL: Zeng and Lin
FP: Follmann and Proschan

1: Henmi and Copas Stat Med. 2010, 2: Brockwell and Gordon Stat Med 2007, 3: Preuß and Ziegler Methods Inf Med. 2014, 4: Zeng and Lin Biometrika. 2015, 5: Huizenga et al Br J Math Stat Psychol. 2011

Comparative evaluation of the methods

vii. Bayesian credible intervals

- ☑ Bayesian intervals produce intervals with coverage closer to the nominal level compared to the HKSJ, modified HKSJ, and PL CIs. [1, 2]
- ☑ Bayesian intervals tend to be smaller than the HKSJ CI even in situations with similar or larger coverage. [1]
- ☒ The performance of the Bayesian intervals may vary depending on the prior assigned to the between-study variance. [3]

Implement in RevMan?	
Bayes	?

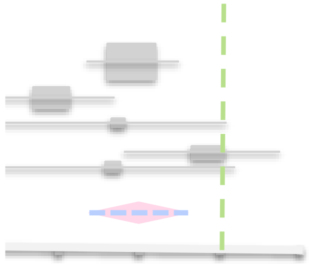
1: Friede et al RSM 2017, 2: Bodnar et al Stat Med. 2017, 3: Lambert et al Stat Med. 2005

HKSJ: Hartung-Knapp/Sidik-Jonkman

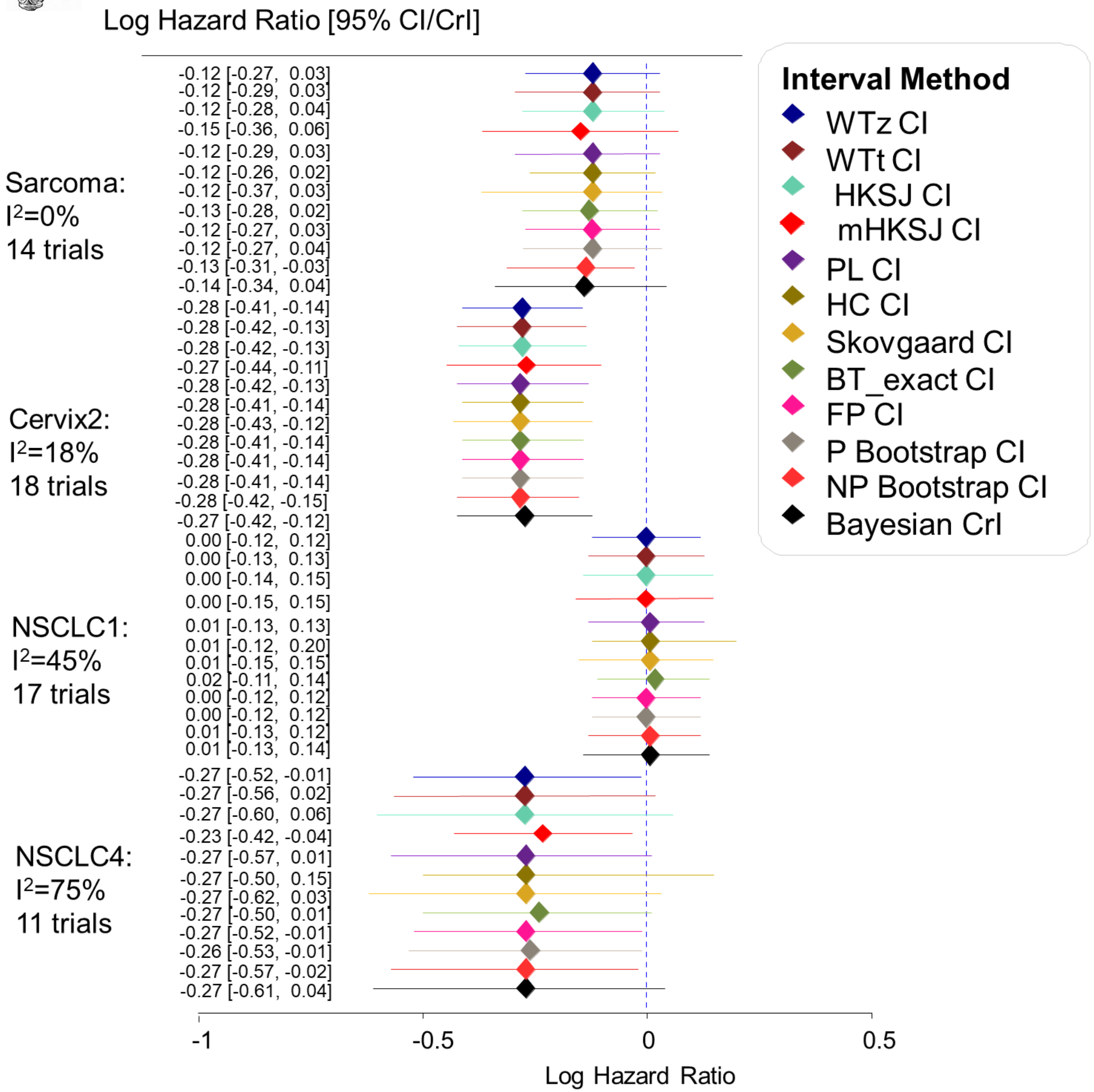
PL: Profile Likelihood

Software for CIs for the overall effect size

CI Method	Software	CI Method	Software	CI Method	Software
WTz	CMA, Excel (MetaEasy, MetaXL), Meta-Disc, Metawin, MIX, MLwin, Open Meta Analyst, RevMan, R, SAS, Stata, SPSS	*PL*	Excel (MetaEasy), HLM, Meta-Disc, MLwin, R, SAS, Stata	*Bootstrap (parametric and non-parametric)*	Metawin, MLwin, R, Stata
WTt	Excel (MetaEasy), R, SAS	*BES*	-	*FP*	Excel (MetaEasy), R, Stata
WTqa	-	*SS*	R	*ZL*	-
HKSJ	CMA, R	*HC*	R	*Bayes*	MLwin, R, SAS, BUGS, OpenBUGS, WinBUGS
Modified HKSJ	Stata	*BT*	R		



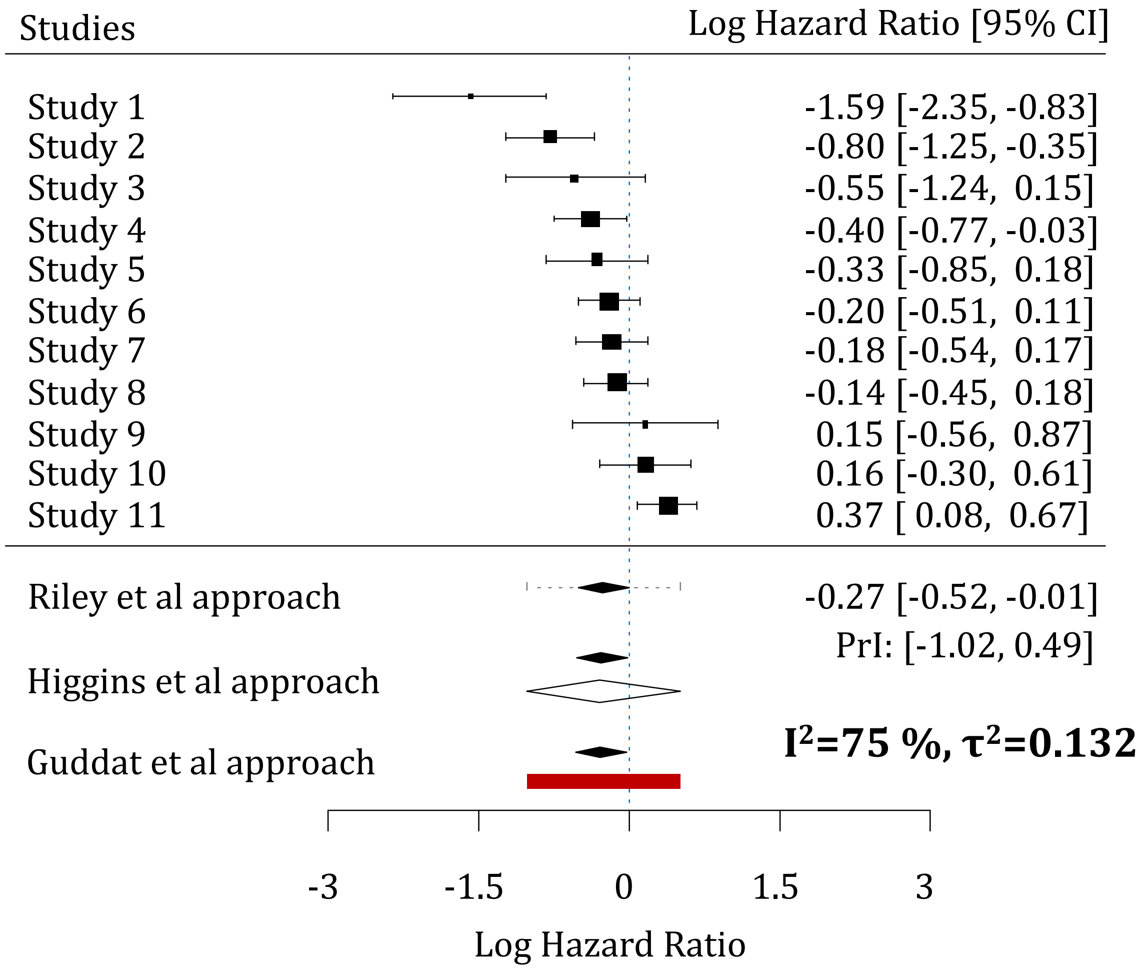
Illustrative example



- The WTz CI lies among the narrowest intervals.
- The Skovgaard statistic CI and the Bayesian CrI lie among the largest intervals.
- For very low (Sarcoma) and low (Cervix2) I^2 values, the modified HKSJ CI has the largest width across all intervals.
- For moderate I^2 value (NSCLC1) the HC CI is associated with the highest uncertainty around the overall effect size.
- For substantial I^2 value (NSCLC4)the HKSJ is the widest CI.

Prediction Interval

- Although prediction intervals have not often been employed in practice they provide useful additional information to the confidence intervals.



- A prediction interval provides a predicted range for the true effect size in a new study:

$$\hat{\mu}_{RE} \pm t_{k-1,0.975}\sqrt{\hat{\tau}^2 + var(\hat{\mu}_{RE})}$$

- Conclusions drawn from a prediction interval are based on the assumption the study-effects are normally distributed

Prediction Interval

- Prediction intervals are particularly helpful when excess heterogeneity exists, and the combination of individual studies into a meta-analysis would not be advisable.

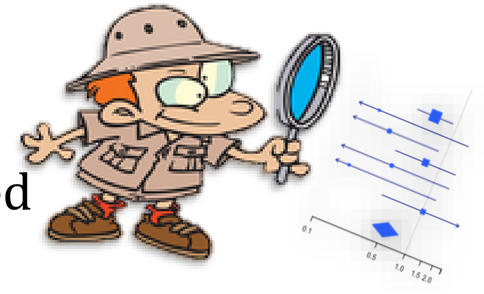
- The 95% prediction interval in >70% of the statistically significant meta-analyses in the Cochrane Database with $\hat{\tau}^2 > 0$, showed that the effect size in a new study could be null or even in the opposite direction from the overall result. [1]

- The 95% prediction interval is only accurate when heterogeneity is large (I^2>30%) and the study sizes are similar. [2]

- For small heterogeneity and different study sizes the coverage of prediction interval can be as low as 78% depending on the between-study variance estimator. [2]

1: IntHout et al BMJ Open 2016, 2: Partlett and Riley Stat Med. 2017

In Summary



- The WTz CI using the DL estimator for the between-study variance, are commonly used and are the default option in many meta-analysis software.

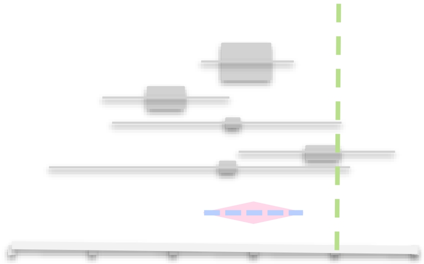
- The accuracy of the WTz CI is not optimal, as coverage can deviate considerably from the nominal level in small meta-analyses.

- Likelihood-based CIs yield coverage closer to the nominal level vs. WTz, but are computationally more demanding than WTz.

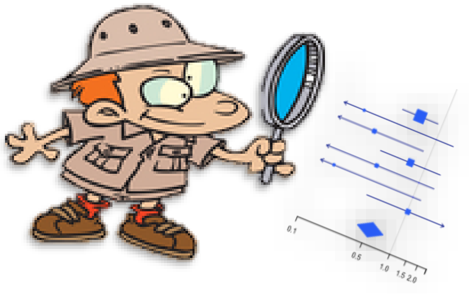
- Overall, studies suggest that the HKSJ method has one of the best performance profiles – performs well even for $k<10$ and is robust across different τ^2 estimators and values.

- But, for $\hat{\tau}^2 = 0$ the HKSJ CI is too narrow. In such cases, the modified HKSJ can be used.

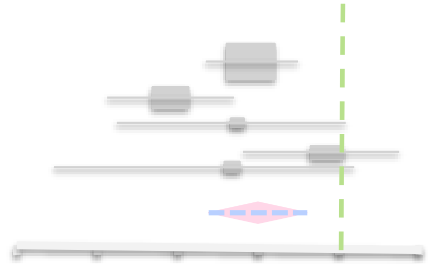
- Caution is also needed in meta-analyses with rare events, with <5 studies, and different study precisions – the modified HKSJ can be used, but not for $k=2$.



In Summary



- The likelihood based methods (SS and BES) have good coverage properties, but have never been compared directly to HKSJ.
- Bayesian intervals may be considered preferable to frequentist intervals in situations where prior information is available.
- The computation of prediction intervals in meta-analysis is valuable. The use of k-1 degrees of freedom rather than k-2 to calculate prediction intervals may be preferable, since the CIs using a t-distribution (e.g., WTt and HKSJ CIs) and prediction intervals will be identical when $\hat{\tau}^2 = 0$.
- We suggest conducting a sensitivity analysis using a variety of methods (with at least 2 to 3 methods) to assess the robustness of findings and conclusions, especially in a meta-analysis with fewer than 10 studies.



References

1. Cornell JE, et al. Random-effects meta-analysis of inconsistent effects: A time for change. *Ann Intern Med.* 2014.
2. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Stat Med.* 2001;20(6):825-840.
3. Bender R, Friede T, Koch A, et al. Methods for evidence synthesis in the case of very few studies. *Res Synth Methods.* 2018;epub ahead of print:1-11.
4. Brockwell SE, Gordon IR. A simple method for inference on an overall effect in meta-analysis. Stat Med. 2007;26(25):4531-4543.
5. Guolo A. Higher-order likelihood inference in meta-analysis and meta-regression. Stat Med. 2012;31(4):313-327.
6. Follmann DA, Proschan MA. Valid inference in random effects meta-analysis. Biometrics. 1999;55(3):732-737.
7. Jackson D, White IR. When should meta-analysis avoid making hidden normality assumptions? Biometrical. 2018.
8. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Stat Med.* 1996;15(6):619-629.
9. Hartung J. An alternative method for meta-analysis. Biom J 1999;41(8):901-916.
10. Hartung J, Knapp G. On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Stat Med*. 2001;20(12):1771-1782.
11. Henmi M, Copas JB. Confidence intervals for random effects meta-analysis and robustness to publication bias. *Stat Med.* 2010;29(29):2969-2983.
12. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc.* 2009;172(1):137-159.
13. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol.* 2014.
14. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. Stat Med. 2003;22(17):2693-2710.
15. Langan D, Higgins JPT, Jackson D, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. Res Synth Methods. 2018;Accepted.
16. Noma H. Confidence intervals for a random-effects meta-analysis based on Bartlett-type corrections. *Stat Med.* 2011.
17. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ.* 2011;342:d549.
18. Sanchez-Meca J, Marin-Martinez F. Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychol Methods.* 2008;13(1):31-48.
19. Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. *Stat Med.* 2002;21(21):3153-3159.

Special thanks to:

- My collaborators: Dr. Dan Jackson, Prof. Ralf Bender, Dr. Oliver Kuss, Dr. Dean Langan, Prof. Julian PT Higgins, Dr. Guido Knapp, Dr. Georgia Salanti
- The European Union's Horizon 2020 (No 754936)

E-mail: averonik@cc.uoi.gr

EVIDENCE SYNTHESIS METHODS STATISTICS TEAM

Visit also our website: https://esm.uoi.gr/