Methods to estimate the between-study variance and to calculate uncertainty in the estimated overall effect size

October 09, 2018

School of Education,
University of Ioannina,
Ioannina,
Greece
Competing Interests

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

- To give an overview of the available methods for estimation of the between-study variance and its corresponding uncertainty
  - Can different methods impact our decision-making?

- To give an overview of the available methods to calculate confidence intervals for the overall effect size
  - What are the properties of the different methods?

- To present real-life and simulation findings that compare the methods
  - Which method is the most appropriate to apply? Are any methods preferable than others?

- To discuss potential issues surrounding the computation of prediction intervals
Work conducted on behalf of the Cochrane Statistical Methods Group

Methods to estimate the between-study variance and its uncertainty in meta-analysis
Areti Angeliki Veroniki, Dan Jackson, Wolfgang Viechtbauer, Ralf Bender, Jack Bowden, Guido Knapp, Oliver Kuss, Julian PT Higgins, Dean Langan, and Georgia Salanti

Recommendations for quantifying the uncertainty in the summary intervention effect and estimating the between-study heterogeneity variance in random-effects meta-analysis
Areti Angeliki Veroniki, Dan Jackson, Wolfgang Viechtbauer, Ralf Bender, Guido Knapp, Oliver Kuss, Dean Langan

Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis
Areti Angeliki Veroniki, Dan Jackson, Ralf Bender, Oliver Kuss, Dean Langan, Julian PT Higgins, Guido Knapp, Georgia Salanti

Inference for the between-study heterogeneity variance
The heterogeneity variance can be estimated using various approaches, including the method proposed by DerSimonian and Laird. This is the most commonly implemented approach.
Work conducted on behalf of the Cochrane Statistical Methods Group

Acknowledgments:

- Dr. Dan Jackson
- Prof. Ralf Bender
- Dr. Oliver Kuss
- Dr. Dean Langan
- Prof. Julian PT Higgins
- Dr. Guido Knapp
- Dr. Jack Bowden
- Dr. Wolfgang Viechtbauer
- Dr. Georgia Salanti
Introduction

- The choice of the method for estimating
  - between-study variance (heterogeneity) and its uncertainty
  - uncertainty for the overall effect size

  is important when conducting a meta-analysis.

- When no appropriate methods are used, this can seriously jeopardize results, and interpretation difficulties may occur.
Have you ever used a different, other than the default option, between-study variance estimator?

a) Yes, I have used different methods in one meta-analysis

b) Yes, I have used different methods in different meta-analyses

c) No, I always use the default option

d) No, I was not aware that different methods exist
### Illustrative example

<table>
<thead>
<tr>
<th>Study (year)</th>
<th>Corticosteroids n/N</th>
<th>Control n/N</th>
<th>Risk ratio (random) (95% CI)</th>
<th>Weight (%)</th>
<th>Risk ratio (random) (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>O’Toole (1969)</td>
<td>6/11</td>
<td>9/12</td>
<td>5.4</td>
<td>5.4</td>
<td>0.73 (0.39 to 1.37)</td>
</tr>
<tr>
<td>Girgis (1991)</td>
<td>72/145</td>
<td>79/135</td>
<td>45.5</td>
<td>45.5</td>
<td>0.85 (0.68 to 1.05)</td>
</tr>
<tr>
<td>Kumarvelu (1994)</td>
<td>5/20</td>
<td>7/21</td>
<td>2.3</td>
<td>2.3</td>
<td>0.75 (0.28 to 1.98)</td>
</tr>
<tr>
<td>Chotmongkol (1996)</td>
<td>5/29</td>
<td>2/30</td>
<td>0.8</td>
<td>0.8</td>
<td>2.59 (0.54 to 12.29)</td>
</tr>
<tr>
<td>Schoeman (1997)</td>
<td>4/67</td>
<td>13/67</td>
<td>1.9</td>
<td>1.9</td>
<td>0.31 (0.11 to 0.90)</td>
</tr>
<tr>
<td>Lardizabal (1998)</td>
<td>4/29</td>
<td>6/29</td>
<td>1.6</td>
<td>1.6</td>
<td>0.67 (0.21 to 2.12)</td>
</tr>
<tr>
<td>Thwaites (2004)</td>
<td>87/274</td>
<td>112/271</td>
<td>42.5</td>
<td>42.5</td>
<td>0.77 (0.61 to 0.96)</td>
</tr>
</tbody>
</table>

**Total DL (95% CI)**

<table>
<thead>
<tr>
<th>0.2</th>
<th>0.5</th>
<th>1</th>
<th>2</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.79</td>
<td>0.69</td>
<td>0.92</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Favors corticosteroids**

**Favors control**

---

**Which is the most appropriate method to use?**

**Pros**

**Cons**

---

*Thorlund et al. RSM 2011*

RE meta-analysis of corticosteroids for preventing death caused by tuberculosis meningitis.
1. Inference on the heterogeneity

\[ \tau^2 \]
Literature Review of the between-study variance methods

Our search identified:

- **16 methods** to estimate the between-study variance (grouped in 5 broad categories).
- **9 methods** to calculate the confidence interval for the between-study variance (grouped in 6 broad categories)

The properties of the methods were evaluated in multiple simulation studies and/or real-life data evaluations of ≥2 methods

---

Select the most appropriate estimator

1. Is a **zero value** possible?
   - Estimators can be **positive** (with solutions **excluding** the zero value) or **non-negative** (with solutions **including** the zero value)

2. Is the estimator **unbiased**?
   
   \[
   \text{Bias}(\hat{\tau}^2) = E(\hat{\tau}^2) - \tau^2 = 0
   \]

3. Is the estimator **efficient**?
   - Low Mean Squared Error (MSE):
     
     \[
     \text{MSE}(\hat{\tau}^2) = E[(\hat{\tau}^2 - \tau^2)^2] = \text{Var}(\hat{\tau}^2) + (\text{Bias}(\hat{\tau}^2))^2
     \]
Select the most appropriate estimator

4. Ease of computation

- Does the method include many and complex steps to estimate heterogeneity?

- Is the method **direct** or **iterative**?

  **Direct methods**: provide an estimator in predetermined number of steps

  **Iterative methods**: converge to a solution when a specific criterion is met

  - **Iterative** methods do not always produce a result because of **failure to converge** during iterations – e.g., ML depends on the choice of maximization method

Be aware of the different **properties** of each estimator!
Method of Moments Estimators

• The method of moments estimators can be categorized to:
  a) Cochran’s Q-based methods

\[ Q = \sum_{i=1}^{k} w_{i,FE} (y_i - \hat{\mu}_{FE})^2 \sim \chi^2_{k-1} \]

b) Generalized Q-based methods

\[ Q_{gen} = \sum_{i=1}^{k} w_{i,RE} (y_i - \hat{\mu}_{RE})^2 \sim \chi^2_{k-1} \]

• The Cochran’s Q-statistic and generalized Q-statistic, belong to the ‘Generalized Cochran between-study variance statistics’:

\[ Q_a = \sum_{i=1}^{k} a_i (y_i - \hat{\mu}_a)^2 \sim \chi^2_{k-1} \]

with \( a_i \) the study weights.

Notation
\( w_i \): weight in study \( i \)
\( y_i \): effect size in study \( i \)
\( \mu \): pooled estimate
\( k \): number of studies in meta-analysis
\( \tau^2 \): heterogeneity
FE: fixed-effect model
RE: random-effects model

DerSimonian and Kacker 2007, Jackson 2013
Method of Moments Estimators

- A method of moments estimator can be derived by equating the expected value of $Q_a$ and its observed value.

- Equating $Q_a$ to its expected value and solving for $\tau^2$ we can obtain the generalised method of moments (GMM) estimator:

  $$\hat{\tau}^2_{GMM} = \max \left\{ 0, \frac{Q_a - \left( \sum a_i v_i - \frac{\sum a_i^2 v_i}{\sum a_i} \right)}{\sum a_i - \frac{\sum a_i^2}{\sum a_i}} \right\}$$

- Each method of moments estimator is a special case of the general class of method of moments estimators with different weights $a_i$.

- Under the assumptions of the RE model, known within-study variances, and before truncation of negative values the generalized method moments estimator is unbiased.
Method of Moments Estimators – Cochran’s Q-based methods

i. DerSimonian and Laird (DL)
   □ The weights used are the inverse of the within-study variances
   ✗ The truncation to zero may lead to biased estimators
   ✓ Performs well with low MSE when $\tau^2$ is small
   ✗ Underestimates true heterogeneity when $\tau^2$ is large and particularly when the number of studies is small

ii. Hedges and Olkin (HO)
   □ The weights used are the inverse of the number of studies
   ✓ Performs well in the presence of substantial $\tau^2$ especially when the number of studies is large
   ✗ But produces large MSE
   ✗ Not widely used and produces large estimates

Method of Moments Estimators – Cochran’s Q-based methods

iii. Hartung and Makambi (HM)
- A modification of DL with weights the inverse of the within-study variances - produces **positive** estimates
- Is more efficient than DL and performs well for meta-analyses with small and large studies
- Estimates **higher** \( \tau^2 \) values compared to DL estimator
- For small to medium study sizes and small \( \tau^2 \) it produces **substantial** positive bias

iv. Hunter and Schmidt (HS)
- A modification of DL with weights the inverse of the within-study variances
- Simple to compute
- Is **more efficient** than DL and HO methods
- The method is associated with **substantial** negative bias

Method of Moments Estimators – Generalized Q-based methods

i. Two-step Dersimonian and Laird (DL2)
   - Uses the RE weights, and decreases bias compared to DL

ii. Two-step Hedges and Olkin (HO2)
   - Uses the RE weights, decreases bias compared to DL and HO

iii. Paule and Mandel (PM)
   - Uses the RE weights and is equivalent to empirical Bayes method.
   - Performs best in terms of bias for both dichotomous and continuous data compared to DL, DL2, HO, REML, and SJ
   - For $\tau^2 = 0$ both DL and PM perform well, but as heterogeneity increases PM approximates $\tau^2$ better compared to DL
   - For mix of small & large studies it may produce higher positive bias than DL, HM, & REML

---

Maximum Likelihood Estimators

i. Maximum Likelihood (ML)

- Although it has a small MSE, it is associated with substantial negative bias as $\tau^2$ increases, the number and size of the included studies is small $^{1,2,3,4}$

ii. Restricted Maximum Likelihood (REML)

- REML is less downwardly biased than DL $^{1,2,5}$
- For dichotomous data, and small $\tau^2$ and number of studies REML tends to have greater MSE than DL, but for continuous data DL and REML have comparable MSEs $^{1,2,5,6}$
- REML is less efficient than ML and HS $^1$
- REML is more efficient with smaller MSE than HO $^1$
- It has relatively low bias and has comparable MSE with HM and DL2 $^7$

An approximate REML estimate is also available yielding almost the same results $^{2,4}$

---

Model error variance estimators

i. Sidik and Jonkman (SJ)

☐ Yields always positive values
☐ Has methodological similarities with PM, but SJ is always positive and non-iterative
☑ Has smaller MSE and substantially smaller bias than DL for large $\tau^2$ and number of studies, and vice versa
☒ Produces larger estimates than the DL method
☒ Large positive bias for small to moderate $\tau^2$ and high MSE

DL: DerSimonian and Laird
PM: Paule and Mandel

Bayes Estimators

i. Bayes Modal (BM)

- Yields always positive values
- When $\tau^2$ is positive BM has very low MSE
- Associated with large bias for small $\tau^2$, especially for few and small studies
- For zero $\tau^2$ it performs worse than DL and REML

ii. Rukhin Bayes (RB)

- For small number of studies, RB with mean prior distribution of $\tau^2$ equal to zero has lower bias than DL

iii. Full Bayesian (FB)

- Allows incorporation of uncertainty in all parameters (including $\tau^2$)
- The choice of prior for $\tau$ is crucial when the number of studies is small

Bootstrap methods

i. Non-parametric bootstrap DL (DLb)

- DLb is associated with lower bias than DL and RB positive when the number of studies is greater than 5.
- DLb performs better than DL in identifying the presence of heterogeneity even for few studies.
- Non-parametric bootstrap methods perform well only for a large number of studies.
- DLb has greater bias compared with DL and this is more profound in small meta-analyses.

*Kontopantelis et al 2013*
## Illustrative example

<table>
<thead>
<tr>
<th>Methodology</th>
<th>$I^2=0%$</th>
<th>$I^2=18%$</th>
<th>$I^2=45%$</th>
<th>$I^2=75%$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of studies in the meta-analysis:</td>
<td>14</td>
<td>18</td>
<td>17</td>
<td>11</td>
</tr>
<tr>
<td>DerSimonian and Laird (DL)</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.13</td>
</tr>
<tr>
<td>Positive DerSimonian and Laird (DLp)</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.13</td>
</tr>
<tr>
<td>Two-step DerSimonian and Laird (DL2)</td>
<td>0.00</td>
<td>0.01</td>
<td>0.04</td>
<td>0.18</td>
</tr>
<tr>
<td>Hedges and Olkin (HO)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
<td>0.22</td>
</tr>
<tr>
<td>Two-step Hedges and Olkin (HO2)</td>
<td>0.00</td>
<td>0.01</td>
<td>0.04</td>
<td>0.19</td>
</tr>
<tr>
<td>Paule and Mandel (PM)</td>
<td>0.00</td>
<td>0.01</td>
<td>0.04</td>
<td>0.19</td>
</tr>
<tr>
<td>Hartung and Makambi (HM)</td>
<td>0.02</td>
<td>0.03</td>
<td>0.06</td>
<td>0.17</td>
</tr>
<tr>
<td>Hunter and Schmidt (HS)</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.11</td>
</tr>
<tr>
<td>Maximum likelihood (ML)</td>
<td>0.00</td>
<td>0.02</td>
<td>0.02</td>
<td>0.13</td>
</tr>
<tr>
<td>Restricted maximum likelihood (REML)</td>
<td>0.00</td>
<td>0.02</td>
<td>0.02</td>
<td>0.16</td>
</tr>
<tr>
<td>Sidik and Jonkman (SJ)</td>
<td>0.07</td>
<td>0.05</td>
<td>0.07</td>
<td>0.21</td>
</tr>
<tr>
<td>Positive Rukhin Bayes (RBp)</td>
<td>0.15</td>
<td>0.11</td>
<td>0.12</td>
<td>0.20</td>
</tr>
<tr>
<td>Full Bayes (FB) [Half normal prior for $\tau$]</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>0.18</td>
</tr>
<tr>
<td>Bayes Modal (BM)</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
<td>0.16</td>
</tr>
<tr>
<td>Non-parametric Bootstrap DerSimonian and Laird (DLb)</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.13</td>
</tr>
</tbody>
</table>

*Veroniki et al. Res Synth Methods 2015*
In summary...

Simulation studies suggest in terms of **bias**:
- DL, DL2, DLp, ML, HS, REML, RB with prior equal to zero, perform well for small $\tau^2$
- HO, HO2, HM, SJ, PM, RBp, BM, perform well for large $\tau^2$

All methods decrease bias as $k$ increases

Simulation studies suggest in terms of **efficiency**:
- DL, ML, HS, REML, perform well for small $\tau^2$
- HO, BM, SJ, PM perform well for large $\tau^2$

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>Zero value included</th>
<th>Simple to compute</th>
<th>Direct</th>
<th>Zero value included</th>
<th>Simple to compute</th>
</tr>
</thead>
<tbody>
<tr>
<td>DL</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>HS</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DLp</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>ML</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>DL2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>REML</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>DLb</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td>AREML</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>HO</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>SJ</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>HO2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>RB</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>PM</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>FB</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>HM</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>BM</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>
## Software for the between-study variance estimator

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>Software</th>
<th>Estimation Method</th>
<th>Software</th>
<th>Estimation Method</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DL</strong></td>
<td>CMA, Excel (MetaEasy), Meta-Disc, Metawin, MIX, Open Meta Analyst, RevMan, R, SAS, STATA, SPSS</td>
<td><strong>ML</strong></td>
<td>CMA, Excel, HLM, Meta-Disc, Metawin, MLwin, Open Meta Analyst, R, SAS, STATA, SPSS</td>
<td><strong>REML</strong></td>
<td>HLM, Meta-Disc, MLwin, Open Meta Analyst, R, SAS, STATA</td>
</tr>
<tr>
<td><strong>HO</strong></td>
<td>R, Open Meta Analyst</td>
<td><strong>PM</strong></td>
<td>Open Meta Analyst, R, SAS, STATA</td>
<td><strong>FB</strong></td>
<td>Mlwin, R, SAS, BUGS, OpenBUGS, WinBUGS</td>
</tr>
<tr>
<td><strong>HM</strong></td>
<td>-</td>
<td><strong>SJ</strong></td>
<td>R, Open Meta Analyst</td>
<td><strong>RB</strong></td>
<td>-</td>
</tr>
<tr>
<td><strong>HS</strong></td>
<td>R</td>
<td><strong>AREML</strong></td>
<td>SPSS</td>
<td><strong>BM</strong></td>
<td>R, STATA</td>
</tr>
<tr>
<td><strong>DL2</strong></td>
<td>-</td>
<td><strong>HO2</strong></td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Which software do you usually prefer to conduct your meta-analyses?

a) Review Manager

b) Stata and/or R

c) WinBUGS/OpenBUGS

d) All of the above

e) None of the above
Should we consider additional options in RevMan?

Which estimation method for the between-study variance should we consider adding in the Cochrane Review Manager?
Recommendations based on published studies

According to simulation and empirical findings, the main factors that may affect the between-study variance estimation are:

- Number and size of studies included in the meta-analysis
- Magnitude of heterogeneity
- Distribution of true treatment effects
- Type of data (e.g., dichotomous, continuous)
- Choice of effect measure
- Frequency of events (for dichotomous outcomes)
- How well study-specific weights, variances and treatment effects are estimated – we often assume these are known.
Recommendations based on published studies

An empirical study using 57,397 Cochrane meta-analyses with $k \geq 2$ showed that:

→ The mean $\tau^2$ is higher than generally assumed but fails to be detected, especially for small $k$!

A descriptive analysis of Cochrane systematic reviews found that 75% of meta-analyses contained 5 or fewer studies

The majority of the pairwise meta-analyses have:

$$k \leq 10 \text{ and } \tau^2 \leq 0.4$$

Summarizing study results in specific scenarios, we make recommendations mostly on NON-Bayesian estimators

• The fully Bayesian estimator has not been evaluated extensively in comparative studies
Recommendations based on published studies

For the most common scenario for pairwise meta-analyses research studies have shown ($k \leq 10$, $\tau^2 \leq 0.4$):

- DL **underestimates** $\tau^2$ when $k$ is small and for rare events 1, 2, 3, 7
- DLp, HM, RBp, BM and SJ **overestimate** $\tau^2$ 2, 4, 5, 6
- DLp has good coverage for the overall effect size 8
- HM has a good coverage for the overall effect size when $\tau^2 \cong 0.07$ for dichotomous outcomes, and for $0.01 \leq \tau^2 \leq 0.05$ for continuous outcomes 8
- DL has **lower bias** and **MSE** than HO and SJ 1, 2
- BM performs **worse** than DL and REML when $\tau = 0$ 3

---

Recommendations based on published studies

“One should probably avoid the biased HS and ML estimators because they can potentially provide quite misleading results”  

- HS and ML are associated with substantial **negative bias**  
- DLb has **higher bias** than DL for small k  
- DLb has good coverage for the overall effect size  
- DL2 approximates PM, inherits most of the best properties of DL and PM and is simple to compute. For rare events underestimates $\tau^2$  
- HO2 approximates PM  
- REML is **less** downwardly biased than DL and ML, but has greater MSE  
  - REML is recommended for continuous data  
  - REML has similar properties with the DL2  
- AREML yields almost identical estimates with REML  

---

<table>
<thead>
<tr>
<th>Implement in RevMan?</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HS</td>
<td>✗</td>
</tr>
<tr>
<td>ML</td>
<td>✗</td>
</tr>
<tr>
<td>DLb</td>
<td>✗</td>
</tr>
<tr>
<td>DL2</td>
<td>?</td>
</tr>
<tr>
<td>HO2</td>
<td>?</td>
</tr>
<tr>
<td>REML</td>
<td>✓</td>
</tr>
<tr>
<td>AREML</td>
<td>✓</td>
</tr>
</tbody>
</table>

---

Recommendations based on published studies

- PM is positively biased when study sizes differ importantly \(^9\)
- it is often approximately unbiased when DL is negatively biased \(^9\)
- PM outperforms DL and REML in terms of bias \(^3, 4, 6, 8\)
  - PM performs better than DL, DL2, PM, HO, REML, SJ in terms of bias for both continuous and dichotomous data \(^7\)
- Easy to obtain
- An improved PM is available for rare events \(^4\)

**BUT**

- Estimation of between-study variance in meta-analyses with <10 studies may be imprecise, especially when study sizes are small and events are rare
- Hence, it is rarely appropriate to rely on one between-study variance estimator!

Implement in RevMan?

| PM   | ✓ |

Confidence Interval (CI) for the between-study variance

Desirable properties

- **Accuracy** = High Coverage Probability – \(P(\tau \in CI)\)
  - 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.
Different CI methods may suggest different results...

![Diagram showing estimated between-study variance ($\hat{\tau}^2$)]

**Notes in the plot**
- The QP, BJ, and Jackson methods used the DL estimator
- The PL and Wt methods used the ML estimator
- The SJ method used the SJ estimator
- The non-parametric bootstrap method used the DLb estimator
- The Bayesian CrI used the FB estimator

Confidence Intervals (CIs) for the between-study variance

- **Bootstrap** CIs have **less** than adequate coverage probabilities

- The **PL** and **Wt** CIs require a **large number of studies** to perform well

- **SJ** has very **poor coverage** probability when $\tau^2$ is small

- **QP** is **preferable** to **PL, Wt, BT and SJ** methods regarding coverage even for a small number of studies

- **QP** and **BJ** provide **narrow** CIs

**Categories**

A. Likelihood-based CIs   
   a) Profile likelihood (PL)

B. Asymptotically normal based CIs
   a) Wald type (Wt)

C. Generalized Cochran Q - based CIs
   a) Biggerstaff and Tweedie (BT)
   b) Jackson (J) (including Biggerstaff and Jackson (BJ))
   c) Q-profile (QP)

D. Sidik and Jonkman CIs (SJ)

E. Bootstrap CIs

F. Bayesian Credible Intervals

---

Confidence Intervals (CIs) for the between-study variance

QP, BJ, and Jackson methods can result in null sets for the CI of $\tau^2$ when heterogeneity and the number of studies are small.

- QP provides more accurate CIs than BJ for large $\tau^2$, and vice versa for small $\tau^2$. For moderate $\tau^2$, Jackson’s method is recommended using weights equal to the reciprocal of the within-study standard errors.$^1, 7$

QP is simple to compute.

Categories

A. Likelihood-based CIs
   a) Profile likelihood (PL)

B. Asymptotically normal based CIs
   a) Wald type (Wt)

C. Generalized Cochran Q-based CIs
   a) Biggerstaff and Tweedie (BT)
   b) Jackson (J) (including Biggerstaff and Jackson (BJ))
   c) Q-profile (QP)

D. Sidik and Jonkman CIs (SJ)

E. Bootstrap CIs

F. Bayesian Credible Intervals

<table>
<thead>
<tr>
<th></th>
<th>PL</th>
<th>Wt</th>
<th>BT, BJ, Jackson</th>
<th>QP</th>
<th>SJ</th>
<th>Bootstrap</th>
<th>Bayesian CRl</th>
</tr>
</thead>
<tbody>
<tr>
<td>DL</td>
<td>--</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>--</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>DLP</td>
<td>--</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>--</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>DL2</td>
<td>--</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>--</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>HO</td>
<td>--</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>--</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>HO2</td>
<td>--</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>--</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>PM</td>
<td>--</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>HM</td>
<td>--</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>HS</td>
<td>--</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>ML</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>--</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>REML</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>--</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>AREML</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>--</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>SJ</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>RB</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>RBp</td>
<td>--</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>FB</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>BM</td>
<td>--</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>DLb</td>
<td>--</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

Not all confidence intervals are appropriate for all of the available between-study variance estimators.
2. Inference on the overall effect size

\[ \mu \]

CI for \( \mu \)
Have you ever used different methods to calculate the uncertainty in the overall effect size?

a) Yes, I have used different methods in one meta-analysis

b) Yes, I have used different methods in different meta-analyses

c) No, I always use the default option

d) No, I was not aware that different methods exist
Various CIs can lead to different conclusions

Which is the most appropriate method to use?

Cornel et al. Annals of Internal Medicine 2014
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 ≥2 methods

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

The most popular technique is WTz

## Confidence Interval methods

<table>
<thead>
<tr>
<th>No</th>
<th>Method</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Wald-type normal distribution (WTz)</td>
<td>$\hat{\mu}<em>{RE} \pm z</em>{0.975}\sqrt{\text{var}(\hat{\mu}_{RE})}$</td>
</tr>
<tr>
<td>2</td>
<td>Wald-type t-distribution (WTt)</td>
<td>$\hat{\mu}<em>{RE} \pm t</em>{k-1,0.975} \sqrt{\text{var}(\hat{\mu}_{RE})}$</td>
</tr>
<tr>
<td>3</td>
<td>Quantile approximation (WTqa)</td>
<td>$\hat{\mu}<em>{RE} \pm b_k \sqrt{\text{var}(\hat{\mu}</em>{RE})}$, with $b_k$ the quantile approximation function of the distribution of the statistic $M = \frac{\hat{\mu}<em>{RE}-\mu}{\sqrt{\text{var}(\hat{\mu}</em>{RE})}}$</td>
</tr>
<tr>
<td>4</td>
<td>Hartung-Knapp/Sidik-Jonkman (HKSJ)</td>
<td>$\hat{\mu}<em>{RE} \pm t</em>{k-1,0.975} \sqrt{\frac{\sigma_{w,\hat{\mu}<em>{RE}}^2}{\hat{\sigma}</em>{w,\hat{\mu}<em>{RE}}^2}}$, with $\sigma</em>{w,\hat{\mu}<em>{RE}}^2 = q \cdot \text{var}(\hat{\mu}</em>{RE})$, $q = \frac{Q_{gen}}{k-1}$, and $Q_{gen} = \sum w_{i,RE}(y_i - \hat{\mu}_{RE})^2$</td>
</tr>
<tr>
<td>5</td>
<td>Modified HKSJ</td>
<td>HKSJ, but use $q^*$ instead of $q$:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$q^* = \max{1, q}$</td>
</tr>
<tr>
<td>6</td>
<td>Profile likelihood (PL)</td>
<td>Profile log-likelihood for $\mu$: $lnL_p(\mu) = lnL(\mu, \hat{t}<em>{ML}^2(\mu))$, $\frac{lnL_p(\mu) - lnL_p(\hat{\mu}</em>{RE})}{\chi^2_{1,0.05}} &gt; 0$</td>
</tr>
</tbody>
</table>
## Confidence Interval methods

<table>
<thead>
<tr>
<th>No</th>
<th>Method</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>7, 8</td>
<td>Higher-order likelihood inference methods</td>
<td>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</td>
</tr>
<tr>
<td>9</td>
<td>Henmi and Copas (HC)</td>
<td><strong>Hybrid approach:</strong> the FE estimate is accompanied by a CI that allows for ( \tau^2 ) under the assumptions of a RE model</td>
</tr>
<tr>
<td>10</td>
<td>Biggerstaff and Tweedie (BT)</td>
<td>[ \hat{\mu}<em>{RE}^{BT} \pm z</em>{0.975} \sqrt{\text{var}(\hat{\mu}<em>{RE}^{BT})}, \quad \text{with } \text{var}(\hat{\mu}</em>{RE}^{BT}) = \frac{1}{(\sum w_{i,RE}^{BT})^2} \sum (w_{i,RE}^{BT})^2 (v_i + \hat{\tau}^2) ] and [ w_{i,RE}^{BT} = E(\hat{w}_{i,RE}) ]</td>
</tr>
<tr>
<td>11</td>
<td>Resampling methods: Zeng and Lin (ZL)</td>
<td><strong>Simulate</strong> values of ( \tau^2 ) using DL, then <strong>simulate</strong> estimated average effect sizes using the sampled ( \tau^2 ) to calculate the weights in ( \hat{\mu}<em>{RE} = \frac{\sum y_i w</em>{i,RE}}{\sum w_{i,RE}} ). Repeat both aspects ( B ) times, get empirical distribution of ( \hat{\mu}_{RE} ) and compute CI</td>
</tr>
<tr>
<td>12, 13</td>
<td>Resampling methods: Bootstrap confidence intervals</td>
<td>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</td>
</tr>
</tbody>
</table>
### Confidence Interval methods

<table>
<thead>
<tr>
<th>No</th>
<th>Method</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>Resampling methods: Follmann and Proschan (FP)</td>
<td><strong>Permutation tests</strong> can be extended to calculate CIs for the effect size. CIs are constructed by <strong>inverting hypothesis test to give the CI bounds</strong> - parameter values that are not rejected by the hypothesis test lie within the corresponding CI</td>
</tr>
<tr>
<td>15</td>
<td>Bayesian credible intervals</td>
<td>Bayesian credible intervals for the overall effect size can be obtained within a Bayesian framework</td>
</tr>
</tbody>
</table>
Should we consider additional options in RevMan?

**Inference on summary effect**

Should we consider adding an extra method to calculate the uncertainty in the overall effect size in the Cochrane Review Manager?
Comparative evaluation of the methods

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

☑️ For large number of studies WTz, WTt, and WTqa perform well

☒ WTz performs worse in terms of coverage for small number of studies (k<16) compared with the PL and the WTt methods

☒ WTz and WTt depend on the number of studies, the $\tau^2$ estimator, and the $\tau^2$ magnitude

☒ Coverage of WTz has been found to be as low as 65% (at 95% nominal level) when $I^2$=90% and k=2,3

☒ Coverage of WTt may be below the 95% nominal level, but it becomes conservative (close to 1) when k is small

☒ WTqa and WTt have on average similar coverage, but WTqa outperforms WTz, PL, and ZL CIs – but it is very conservative

☒ WTqa has been criticized that it is very difficult to obtain suitable critical values $b_k$ that apply to all meta-analyses

<table>
<thead>
<tr>
<th>Implement in RevMan?</th>
<th>WTz</th>
<th>WTt</th>
<th>WTqa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Implemented</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>

WTz: Wald type – normal distr

WTt: Wald type – t distr

WTqa: Wald type – quantile approximation

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 \(\tau^2\), 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 \(^6\)

---

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 $\tau^2$ the modified HKSJ tends to be over-conservative ¹,²,³

<table>
<thead>
<tr>
<th>Implement in RevMan?</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HKSJ</td>
<td>✓</td>
</tr>
<tr>
<td>mHKSJ</td>
<td>✓</td>
</tr>
</tbody>
</table>

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 $\tau^2$ 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 $k\leq5$ as **BES** tends to be over-conservative \(^6\)

<table>
<thead>
<tr>
<th>Implement in RevMan?</th>
</tr>
</thead>
<tbody>
<tr>
<td>PL</td>
</tr>
<tr>
<td>BES</td>
</tr>
<tr>
<td>SS</td>
</tr>
</tbody>
</table>

**WTz**: Wald type – normal distr

**WTt**: Wald type – t distr

**PL**: Profile Likelihood

**BES**: Bartlett-type adjusted efficient score statistic

**SS**: Skovgaard's statistic
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?

<table>
<thead>
<tr>
<th></th>
<th>HC</th>
<th>BT</th>
<th>ZL</th>
<th>FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implement?</td>
<td>X</td>
<td>X</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

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

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 \(^3\).

---

# Software for CIs for the overall effect size

<table>
<thead>
<tr>
<th>CI Method</th>
<th>Software</th>
<th>CI Method</th>
<th>Software</th>
<th>CI Method</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>WTz</td>
<td>CMA, Excel (MetaEasy, MetaXL), MetaDisc, Metawin, MIX, MLwin, OpenMeta Analyst, RevMan, R, SAS, Stata, SPSS</td>
<td>PL</td>
<td>Excel (MetaEasy), HLM, MetaDisc, MLwin, R, SAS, Stata</td>
<td>Bootstrap (parametric and non-parametric)</td>
<td>Metawin, MLwin, R, Stata</td>
</tr>
<tr>
<td>WTt</td>
<td>Excel (MetaEasy), R, SAS</td>
<td>BES</td>
<td>-</td>
<td>FP</td>
<td>Excel (MetaEasy), R, Stata</td>
</tr>
<tr>
<td>WTqa</td>
<td>-</td>
<td>SS</td>
<td>R</td>
<td>ZL</td>
<td>-</td>
</tr>
<tr>
<td>HKSJ</td>
<td>CMA, R</td>
<td>HC</td>
<td>R</td>
<td>Bayes</td>
<td>MLwin, R, SAS, BUGS, OpenBUGS, WinBUGS</td>
</tr>
<tr>
<td>Modified HKSJ</td>
<td>Stata</td>
<td>BT</td>
<td>R</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Studies</th>
<th>Log Hazard Ratio [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>-1.59 [-2.35, -0.83]</td>
</tr>
<tr>
<td>Study 2</td>
<td>-0.80 [-1.25, -0.35]</td>
</tr>
<tr>
<td>Study 3</td>
<td>-0.55 [-1.24, 0.15]</td>
</tr>
<tr>
<td>Study 4</td>
<td>-0.40 [-0.77, -0.03]</td>
</tr>
<tr>
<td>Study 5</td>
<td>-0.33 [-0.85, 0.18]</td>
</tr>
<tr>
<td>Study 6</td>
<td>-0.20 [-0.51, 0.11]</td>
</tr>
<tr>
<td>Study 7</td>
<td>-0.18 [-0.54, 0.17]</td>
</tr>
<tr>
<td>Study 8</td>
<td>-0.14 [-0.45, 0.18]</td>
</tr>
<tr>
<td>Study 9</td>
<td>0.15 [-0.56, 0.87]</td>
</tr>
<tr>
<td>Study 10</td>
<td>0.16 [-0.30, 0.61]</td>
</tr>
<tr>
<td>Study 11</td>
<td>0.37 [ 0.08, 0.67]</td>
</tr>
</tbody>
</table>

- 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.  

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

- 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.

Should we consider more **between-study variance estimators** in Review Manager?

a) No because research has not concluded which one is the best

b) Yes because research has not concluded which one is the best

c) No because differences are negligible

d) Yes because results are sensitive
Should we consider more CI methods for the overall effect size in Review Manager?

a) No because research has not concluded which one is the best

b) Yes because research has not concluded which one is the best

c) No because differences are negligible

d) Yes because results are sensitive
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.

- Simulations suggest that PM and REML estimators are better alternatives to estimate the between-study variance than DL.

- The QP method and the alternative approach based on a ‘generalized Cochrane between-study variance statistic’ are among the best options to compute CI around the between-study variance.

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

- 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 $\tau^2$ 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.

- Bayesian methods may be considered preferable when prior information is available.

- A sensitivity analysis using a variety of methods may be needed, particularly when studies are few in number.
References

References


Thank you for your attention!

Questions?

Areti Angeliki Veroniki MSc, PhD
Research Fellow, University of Ioannina, Ioannina, Greece
Post-doctoral Fellow, Faculty of Medicine, Imperial College, London, UK
Affiliate Scientist, St. Michael’s Hospital, Toronto, Canada
E-mail: averonik@cc.uoi.gr