An Introduction to Meta-analysis

Geoff Der
Statistician
MRC Social and Public Health Sciences Unit
In this session:

What is meta-analysis?

When is it appropriate to use?

Statistical methods

Software programmes

Publishing meta-analyses
Karl Pearson (1904) conducted the first meta-analysis commissioned by the British government on the effects of a typhoid vaccination

Gene Glass (1974) coined ‘meta-analysis’:
“…the analysis of analyses. It connotes a rigorous alternative to the casual, narrative discussions of research studies which typify our attempts to make sense of the rapidly expanding research literature”.
What is meta-analysis?

“Statistical combination of results from two or more separate studies” to answer a common question

Why?
To provide a test with more power than separate studies
To summarise numerous and inconsistent findings
To investigate consistency of effect across different samples

http://www.cochrane-handbook.org/
What questions are addressed?

1. What is the direction of the effect?

2. What is the size of the effect?

3. Is the effect consistent across studies? (heterogeneity)

4. What is the strength of evidence for the effect? (quality assessment)

http://www.cochrane-handbook.org/
Some Background – Clinical Trials

Early trials show larger effects than later trials

Better designed trials show smaller effects

Larger trials show smaller effects

‘Natural history’ of novel interventions

Proliferation of small underpowered trials

When is it appropriate?

Observational and Intervention studies

How many studies make it worth while?

Are there additional exclusion criteria for meta-analyses?

- Duplicate publications, e.g. in longitudinal studies
- Very small studies
- Poor quality
- Results not in suitable format? (But can approach authors)

Greenland, Epidemiologic Reviews 1987;9
Statistical Issues

Effect size measures
transformations; direction and magnitude of effect

Heterogeneity: random and fixed effects

Publication bias

Quality assessment and sensitivity analyses:
bias and confounding; subgroup analysis or meta-regression?
Effect Size measure

- A statistic that summarises the observed intervention effect

examples

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Predictor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary</td>
<td>Continuous</td>
</tr>
<tr>
<td>OR RR</td>
<td>SMD</td>
</tr>
<tr>
<td>OR RR per unit</td>
<td>Correlation</td>
</tr>
</tbody>
</table>

OR Odds ratio
RR Relative Risk
SMD standardized mean difference
Effect Size Measures

Standardised Mean difference

- Cohen’s $d$
- Hedges’ $g$
- Glass’s $\Delta$

Binary outcome
- Odds Ratio
- Relative Risk

Survival
- Hazard ratio

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s},$$
Transforming Effect Size Measures

Transform reported effect sizes to common measure
Eg measures of spread/variance: CI, SD, SE, IQR

Converting odds ratios to continuous outcome effect sizes, or vice versa (Chinn, *Statistics in Medicine*, 2000;19:3127)

HR ~ OR ~ RR when the risk of an event is low: <20%

Take care and check results!

Online effect size calculator:
http://www.campbellcollaboration.org/resources/effect_size_input.php

R package ‘compute.es’
Effect Sizes – an example

ES = cohen’s $d$ (eg RCT continuous outcome)

<table>
<thead>
<tr>
<th>Paper reports:</th>
<th>Verdict</th>
<th>Check/Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean, SD, N for each group</td>
<td>ideal</td>
<td></td>
</tr>
<tr>
<td>Effect Size</td>
<td>ideal ?</td>
<td>Is it $d$ ?</td>
</tr>
<tr>
<td>Mean difference + 95% CI (or SE)</td>
<td>excellent</td>
<td>Transform</td>
</tr>
<tr>
<td>Mean difference + SD</td>
<td>good</td>
<td>Check SD</td>
</tr>
<tr>
<td>Mean difference + $p$ value</td>
<td>OK ?</td>
<td>Precision ?</td>
</tr>
<tr>
<td>$t$ value + $p$ value</td>
<td>OK ?</td>
<td>Precision ?</td>
</tr>
<tr>
<td>Median difference</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>Correlation</td>
<td>X</td>
<td>Unlikely</td>
</tr>
</tbody>
</table>
Main output of MA - a Forest Plot

Smith et al. 1991  
Jones et al. 1993  
Smith et al. 1999  
Ng et al. 2004  
Chu et al. 2009  

Summary measure

<table>
<thead>
<tr>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smith et al. 1991</td>
</tr>
<tr>
<td>Jones et al. 1993</td>
</tr>
<tr>
<td>Smith et al. 1999</td>
</tr>
<tr>
<td>Ng et al. 2004</td>
</tr>
<tr>
<td>Chu et al. 2009</td>
</tr>
<tr>
<td><strong>Summary measure</strong></td>
</tr>
</tbody>
</table>
Example: RR with CIs

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Weight</th>
<th>IV, Random, 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ho et al., (M)</td>
<td>3.0%</td>
<td>0.53 [0.25, 1.12]</td>
</tr>
<tr>
<td>Ho et al., (F)</td>
<td>5.9%</td>
<td>0.53 [0.32, 0.87]</td>
</tr>
<tr>
<td>Laurin et al., (M)</td>
<td>6.4%</td>
<td>0.84 [0.53, 1.34]</td>
</tr>
<tr>
<td>Laurin et al., (F)</td>
<td>7.6%</td>
<td>0.55 [0.36, 0.83]</td>
</tr>
<tr>
<td>Schuit et al.,</td>
<td>1.5%</td>
<td>0.56 [0.19, 1.66]</td>
</tr>
<tr>
<td>Yaffe et al.,</td>
<td>16.0%</td>
<td>0.78 [0.64, 0.96]</td>
</tr>
<tr>
<td>Pignatti et al.,</td>
<td>1.4%</td>
<td>0.27 [0.09, 0.82]</td>
</tr>
<tr>
<td>Lytle et al.,</td>
<td>6.4%</td>
<td>0.63 [0.40, 1.00]</td>
</tr>
<tr>
<td>Flicker et al.,</td>
<td>3.4%</td>
<td>0.50 [0.25, 1.00]</td>
</tr>
<tr>
<td>Singh-Manoux et al.,</td>
<td>14.8%</td>
<td>0.81 [0.65, 1.01]</td>
</tr>
<tr>
<td>Sumic et al., (M)</td>
<td>1.1%</td>
<td>0.91 [0.25, 3.36]</td>
</tr>
<tr>
<td>Sumic et al., (F)</td>
<td>1.1%</td>
<td>0.12 [0.03, 0.44]</td>
</tr>
<tr>
<td>Middleton et al.,</td>
<td>15.2%</td>
<td>0.73 [0.59, 0.91]</td>
</tr>
<tr>
<td>Niti et al.,</td>
<td>12.1%</td>
<td>0.60 [0.45, 0.80]</td>
</tr>
<tr>
<td>Etgen et al.,</td>
<td>4.1%</td>
<td>0.44 [0.24, 0.82]</td>
</tr>
</tbody>
</table>

Total (95% CI) 100.0% 0.65 [0.57, 0.75]

Heterogeneity: $\tau^2 = 0.02$; $\chi^2 = 20.93$, df = 14 ($P = 0.10$); $I^2 = 33$

Test for overall effect: $Z = 6.17$ ($P < 0.00001$)

High physical activity & Cognitive decline (Sofi et al, J Internal Med, 2010;269:107-117)
Example: HR with CIs

Childhood IQ and risk of mortality (Calvin et al., 2010)
**Forest Plot— mean difference**

<table>
<thead>
<tr>
<th>Study name</th>
<th>Statistics for each study</th>
<th>N</th>
<th>Age</th>
<th>IQ test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hay</td>
<td>7.92 (2.94)</td>
<td>130</td>
<td>11</td>
<td>WISC</td>
</tr>
<tr>
<td>Ghys</td>
<td>0.21 (2.82)</td>
<td>108</td>
<td>4</td>
<td>GOS</td>
</tr>
<tr>
<td>Morrow-Tlucak</td>
<td>5.83 (2.57)</td>
<td>219</td>
<td>2</td>
<td>MDI</td>
</tr>
<tr>
<td>Gomez-Sanchiz</td>
<td>3.98 (1.53)</td>
<td>164</td>
<td>2</td>
<td>MDI</td>
</tr>
<tr>
<td>Jacobson</td>
<td>4.00 (1.49)</td>
<td>279</td>
<td>11</td>
<td>WISC</td>
</tr>
<tr>
<td>Wigg</td>
<td>0.80 (1.37)</td>
<td>343</td>
<td>12</td>
<td>WISC</td>
</tr>
<tr>
<td>Fergusson</td>
<td>2.09 (0.65)</td>
<td>954</td>
<td>7</td>
<td>WISC</td>
</tr>
<tr>
<td>Richards</td>
<td>0.98 (0.61)</td>
<td>511</td>
<td>8</td>
<td>SC</td>
</tr>
<tr>
<td>NLSY</td>
<td>0.52 (0.36)</td>
<td>5475</td>
<td>10</td>
<td>PIAT</td>
</tr>
</tbody>
</table>

**Difference in means (95% CI)**

- Disadvantage with breast feeding
- Advantage with breast feeding
Example: standardised mean difference

Figure 1  Standardized mean differences in cognitive function test scores (and 95% confidence intervals) between folic acid, with or without other B vitamins, and placebo groups for all tests within each randomized trial and pooled effect across all trials. CI = confidence interval; SMD = standardized mean difference.

Fixed vs Random effects

Fixed Effects
Each study is estimating the same quantity

(methods: Mantel-Haenszel, Peto odds ratio, Inverse variance)

Random Effects
Differences in study sample, design, measurement etc contribute to the effect size

DerSimonian and Laird method
Heterogeneity

Variability between studies caused by differences in:
- Study samples (e.g. healthy, clinical)
- Interventions or outcomes
- Methodology: design, measures, quality etc.

“Statistical heterogeneity manifests itself in the… [study] effects being more different from each other than one would expect due to random error (chance) alone” (Cochrane Handbook)
Assessing heterogeneity

• Visual inspection:
  - confidence intervals have poor overlap
• Formal test:
  - Chi-squared: are observed differences compatible with chance alone?
  (NB. low power with small number of studies; $p > 0.10$ gives greater confidence of no heterogeneity)

• Additionally, look at the impact of heterogeneity on your aggregate estimate: inconsistency ($I^2 > 50\%$)

But, isn’t there always clinical and methodological diversity?
Dealing with heterogeneity

• Check data!
• Choose random effect meta-analysis
• Explore the causes of heterogeneity: subgroup analysis or meta-regression
• Change the effect measure
• Exclude outlying studies
• Consider whether a meta-analysis is the right course

Must be dealt with sensitively and with a good rationale for the methods used
Subgroup Analyses

Dividing your studies by a design feature:
- Participant characteristic (sex, age, clinical diagnoses, geographical region)
- Study design characteristic (type of intervention, length of follow-up, type of measure used, e.g. cognitive function)

NB. More subgroup analyses increase the risk of false negatives and false positives (patients being denied an effective treatment, or given a harmful / ineffective one)
Meta-regression

Linear regression of the effect estimates on some study characteristic

Outcome: Study effect size

Explanatory variable: a characteristic of the studies that may influence the magnitude of the effect (potential effect modifier or covariate)

Regression is weighted by study size/precision
Subgroup analyses & meta-regression
Considerations of both

• Are there enough studies that include the specified characteristics to justify these methods?

• Specify the characteristics in advance

• Keep numbers of characteristics to a minimum

• Is there adequate scientific rationale?

• Does one characteristic confound another?
Meta-regression example
Publication bias / small study bias: Addressing file drawer effects

“To control resulting overall effect sizes for publication bias, several tests were performed. These tests consisted of visual inspection of funnel plots (Light & Pillemer, 1984), Rosenthal's Fail-safe N (Rosenthal, 1979), a weighted Failsafe N (Rosenberg, 2005), Orwin's Fail-safe N (Orwin, 1983), Begg and Mazumdar's rank correlation method (Begg & Mazumdar, 1994), Egger's regression test (Egger, Smith, Schneider, & Minder, 1997; Sterne & Egger, 2005), trim-and-fill analysis (Duval & Tweedie, 2000) following the approach as suggested by Peters, Sutton, Jones, Abrams and Rushton (2007), a sensitivity analysis for publications bias as suggested by Vevea and Woods (2005), and a method based on truncated normal distributions (Formann, 2008).

Application of this multitude of differential approaches originates in the increased awareness of problems of publication bias in general and the corresponding recent developments of enhanced methods to account for it.”

Funnel plots
Funnel plots
Publication bias: Example 1
cognitive epidemiology

Childhood IQ and risk of mortality (Calvin et al., 2010)
Trim-and-fill

1 trim off the asymmetric part of the funnel

2 use the symmetric remainder to estimate the true centre

3 replace the trimmed studies and their missing counterparts

4 estimate the true mean and its variance from the filled funnel plot

Duval & Tweedie, Biometrics, 2000;56(2):455-63.
Quality assessment

To control for bias, particularly in observational studies

Use a quality checklist/tool:
(eg Moher, 1995 for RCTs ; Sanderson, 2007 for observational studies)

Independent quality scoring (and blinded, if poss)

1. Forest Plot ordered by quality score.
   Is there an association?
2. Quality score as in meta-regression
3. Exclude low quality studies
Software: specially built programmes

Comprehensive Meta-Analysis (CMA)
MetAnalysis
MetaWin
MIX - Free
RevMan - Free
WEasyMA

<table>
<thead>
<tr>
<th>Items and subgroups</th>
<th>MIX</th>
<th>CMA</th>
<th>MetaWin</th>
<th>RevMan</th>
<th>WEasyMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>All researchers (26)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall rating (min-max)</td>
<td>8.6 (6.7 to 10)</td>
<td>6.9 (3.7 to 9.7)</td>
<td>6.2 (4.3 to 8.7)</td>
<td>6.1 (4.3 to 8.3)</td>
<td>4.2 (1 to 7.3)</td>
</tr>
<tr>
<td>Getting started</td>
<td>8.6</td>
<td>7.4</td>
<td>6.8</td>
<td>7.6</td>
<td>4.3</td>
</tr>
<tr>
<td>Data preparation</td>
<td>8.3</td>
<td>6.1</td>
<td>6.3</td>
<td>4.5</td>
<td>2.6</td>
</tr>
<tr>
<td>Usability in analysis</td>
<td>8.8</td>
<td>7.1</td>
<td>5.5</td>
<td>6.3</td>
<td>5.9</td>
</tr>
<tr>
<td>Experienced (7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall rating (min-max)</td>
<td>8.1 (7.0 to 9.7)</td>
<td>6.8 (6.0 to 7.2)</td>
<td>5.9 (4.3 to 7.7)</td>
<td>5.4 (4.3 to 6.3)</td>
<td>3.3 (1 to 5.7)</td>
</tr>
<tr>
<td>Getting started</td>
<td>8.0</td>
<td>7.6</td>
<td>6.2</td>
<td>7.5</td>
<td>2.9</td>
</tr>
<tr>
<td>Data preparation</td>
<td>8.3</td>
<td>6.3</td>
<td>6.3</td>
<td>3.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Usability in analysis</td>
<td>8.0</td>
<td>6.6</td>
<td>5.4</td>
<td>6.3</td>
<td>5.3</td>
</tr>
<tr>
<td>Inexperienced (19)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall rating (min-max)</td>
<td>8.7 (6.7 to 10)</td>
<td>7 (3.7 to 9.7)</td>
<td>6.1 (4.3 to 8.7)</td>
<td>6.1 (4.7 to 8.3)</td>
<td>4.6 (1.3 to 7.1)</td>
</tr>
<tr>
<td>Getting started</td>
<td>8.0</td>
<td>7.3</td>
<td>6.9</td>
<td>7.7</td>
<td>5.0</td>
</tr>
<tr>
<td>Data preparation</td>
<td>8.3</td>
<td>6.3</td>
<td>6.3</td>
<td>5.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Usability in analysis</td>
<td>9.1</td>
<td>7.3</td>
<td>5.5</td>
<td>6.3</td>
<td>6.1</td>
</tr>
</tbody>
</table>

Comprehensive Meta-Analysis

The program shows the effect size and confidence interval for each study.

The program shows the relative weight assigned to each study using fixed and random effects.

Select the computational model.

The program shows the combined effect size and confidence interval using fixed and random effects.

<table>
<thead>
<tr>
<th>Model</th>
<th>Study name</th>
<th>Statistics for each study</th>
<th>Weight (Fixed)</th>
<th>Weight (Random)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Risk ratio</td>
<td>Lower limit</td>
<td>Upper limit</td>
</tr>
<tr>
<td>Fixed</td>
<td></td>
<td>0.411</td>
<td>0.134</td>
<td>1.257</td>
</tr>
<tr>
<td>Random</td>
<td></td>
<td>0.205</td>
<td>0.086</td>
<td>0.486</td>
</tr>
<tr>
<td>Both</td>
<td></td>
<td>0.237</td>
<td>0.179</td>
<td>0.312</td>
</tr>
</tbody>
</table>

Select the computational model.
Standard statistical software

R

http://cran.r-project.org/web/packages/rmeta/rmeta.pdf

STATA


SAS

http://www.senns.demon.co.uk/SAS%20Macros/SASMacros.html

WinBUGS (Bayesian)

http://www.openbugs.info/w/
Publishing a meta-analysis

• Consider which journals have an interest in publishing meta-analyses - what are their instructions to authors?

• Does the quantitative reporting of results from meta-analysis reduce the need for qualitative discussion more typical of a systematic review?

• Are there standard protocols for writing up? Yes, MOOSE…

MOOSE Checklist

Reporting of background should include
- Problem definition
- Hypothesis statement
- Description of study outcome(s)
- Type of exposure or intervention used
- Type of study designs used
- Study population

Reporting of search strategy should include
- Qualifications of searchers (e.g., librarians and investigators)
- Search strategy, including time period included in the synthesis and keywords
- Effort to include all available studies, including contact with authors
- Databases and registries searched
- Search software used, name and version, including special features used (e.g., explosion)
- Use of hand searching (e.g., reference lists of obtained articles)
- List of citations located and those excluded, including justification
- Method of addressing articles published in languages other than English
- Method of handling abstracts and unpublished studies
- Description of any contact with authors
MOOSE Checklist cont…

Reporting of methods should include
- Description of relevance or appropriateness of studies assembled for assessing the hypothesis to be tested
- Rationale for the selection and coding of data (e.g., sound clinical principles or convenience)
- Documentation of how data were classified and coded (e.g., multiple raters, blinding, and interrater reliability)
- Assessment of confounding (e.g., comparability of cases and controls in studies where appropriate)
- Assessment of study quality, including blinding of quality assessors; stratification or regression on possible predictors of study results
- Assessment of heterogeneity
- Description of statistical methods (e.g., complete description of fixed or random effects models, justification of whether the chosen models account for predictors of study results, dose-response models, or cumulative meta-analysis) in sufficient detail to be replicated
- Provision of appropriate tables and graphics
MOOSE Checklist cont...

Reporting of results should include
• Graphic summarizing individual study estimates and overall estimate
• Table giving descriptive information for each study included
• Results of sensitivity testing (eg, subgroup analysis)
• Indication of statistical uncertainty of findings

Reporting of discussion should include
• Quantitative assessment of bias (eg, publication bias)
• Justification for exclusion (eg, exclusion of non–English-language citations)
• Assessment of quality of included studies

Reporting of conclusions should include
• Consideration of alternative explanations for observed results
• Generalization of the conclusions (ie, appropriate for the data presented and within the domain of the literature review)
• Guidelines for future research
• Disclosure of funding source
Resources
Introduction to meta-analysis


- Chapter 16 on Meta-analysis of Epidemiological and Observational Studies

Resources
Meta-analytic methods


- Chapters 3 to 9
Resources

Reporting a meta-analysis


MOOSE (Meta-Analysis of Observational Studies in Epidemiology). This checklist for reporting observational studies was developed following a workshop convened to address the problem of increasing diversity and variability that exist in reporting meta-analyses of observational studies. (Stroup et al., 2000). Checklist: http://jama.ama-assn.org/cgi/content/full/283/15/2008/TABLEJST00003T1


- Chapter 10 Reporting the Results of Meta-analysis
Acknowledgements

Thanks to Catherine Calvin who helped prepare version 1 of this presentation

Contact: Geoff.Der@glasgow.ac.uk