An introduction to meta-analysis:
History, methods, misconceptions,
and recent developments

Wolfgang Viechtbauer
Maastricht University
http://www.wvbauer.com

The Information Explosion
• 2,300 biomedical journals in 1940
• now there are close to 25,000
• approximately 27,000+ RCTs per year
• similar growth in other disciplines
• rough estimates:
  • # of articles double every ~10 years
  • # of journals double every ~15 years

1. finding relevant literature
2. accessing the literature
3. maintaining awareness of the literature
4. reading and processing the information

The individual scientist is being
overloaded with scientific
information [...] and can no longer
keep up with and assimilate all
the information being produced
Garvey & Griffith (1971)

How to Summarize the Results?
• traditionally:
  • narrative literature reviews
  • vote counting methods
  • combining tests of significance
• now:
  • systematic reviews
  • meta-analysis
Narrative Literature Review

• a description/summary of the current state of knowledge on a particular topic supported by empirical findings as well as the underlying theories and models
• possible problems:
  • unsystematic
  • subjective
  • intractable
  • in essence scientifically unsound

Vote Counting

• examine all relevant studies conducted
• categorize based on statistical significance
  • statistically significant (with \( T_{rt} > Ctrl \) or \( r > 0 \))
  • not statistically significant
  • statistically significant (with \( T_{rt} < Ctrl \) or \( r < 0 \))
• declare most frequent category the ‘winner’
• inconsistent when power of studies is low (Hedges & Olkin, 1985): as \( k \to \infty \), method fails to find a true effect or association

Combining Tests of Significance

• long history of methods for combining the results from independent significance tests (Tippett, 1931; Fisher, 1932; Pearson, 1933; Stouffer et al., 1949; Wilkinson, 1951; Mosteller & Bush, 1954; Good, 1955; Lipták, 1958; Lancaster, 1961; …)

Fisher’s Method

• if \( H_0 \) is true, then \( p \sim \text{Uniform}(0, 1) \)
• then we can show that \(-2 \ln[p] \sim \chi^2_2\)
• now assume \( H_0 \) is true for \( i = 1, ..., k \) tests
• then \(-2 \sum \ln[p_i] \sim \chi^2_k\)

Example
• want to know if $x$ and $y$ are correlated
• test $H_0: \rho = 0$ in three different studies
• $n = 50$ in all three studies
• find $r_1 = .33, r_2 = .25$, and $r_3 = .15$
• then $p_1 = .02, p_2 = .08$, and $p_3 = .30$
• so $-2\sum \ln(p_i) = 15.28$
• under a $\chi^2$ distribution with $df = 6$, this yields a combined $p$-value of .018
• reject $H_0; \rho_1 = \rho_2 = \rho_3 = 0$

Combining Tests of Significance
• tests a fairly uninteresting null hypothesis
• uses little information from the studies
• also:

\[
\text{Test of Significance} = \frac{\text{Size of the Effect or Strength of Association}}{\text{Size of Study}}
\]

this is what vote counting and combined tests of significance are based on
this is what we typically want to know

History/Origins of Meta-Analysis
• nice summary in Chalmers et al. (2002)
• Pearson (1904) – the earliest MA?
• Cochran et al. work in agriculture
• physics (Birge, 1932)
• origin of term “meta-analysis” (Glass, 1976)
• some early MAs in psychology
• Cochrane and the EBM movement

Pearson (1904)
• maybe the earliest MA
• examined effectiveness of a vaccine against typhoid based on studies conducted among soldiers in the British Empire
• outcomes: infection and mortality
• data summarized in terms of 2 $\times$ 2 tables

### Appendix B.

<table>
<thead>
<tr>
<th></th>
<th>Incubated.</th>
<th>Non-Incubated</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Hospital Staff in South Africa.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exposed</td>
<td>49</td>
<td>49</td>
<td>98</td>
</tr>
<tr>
<td>Cases</td>
<td>70</td>
<td>65</td>
<td>135</td>
</tr>
<tr>
<td>Totals</td>
<td>119</td>
<td>114</td>
<td>233</td>
</tr>
<tr>
<td><strong>2. Garrison of Ladysmith in South Africa.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exposed</td>
<td>23</td>
<td>26</td>
<td>49</td>
</tr>
<tr>
<td>Cases</td>
<td>34</td>
<td>31</td>
<td>65</td>
</tr>
<tr>
<td>Totals</td>
<td>57</td>
<td>57</td>
<td>114</td>
</tr>
<tr>
<td><strong>3. Methven’s Column in South Africa.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exposed</td>
<td>98</td>
<td>98</td>
<td>196</td>
</tr>
<tr>
<td>Cases</td>
<td>102</td>
<td>102</td>
<td>204</td>
</tr>
<tr>
<td>Totals</td>
<td>196</td>
<td>196</td>
<td>392</td>
</tr>
<tr>
<td><strong>4. Single Regiments in South Africa.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exposed</td>
<td>29</td>
<td>29</td>
<td>58</td>
</tr>
<tr>
<td>Cases</td>
<td>30</td>
<td>30</td>
<td>60</td>
</tr>
<tr>
<td>Totals</td>
<td>59</td>
<td>59</td>
<td>118</td>
</tr>
<tr>
<td><strong>5. Army in India.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exposed</td>
<td>10</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Cases</td>
<td>11</td>
<td>11</td>
<td>22</td>
</tr>
<tr>
<td>Totals</td>
<td>21</td>
<td>21</td>
<td>42</td>
</tr>
</tbody>
</table>

The following table gives the results of calculating the correlation coefficients of the tables in Appendix B:

<table>
<thead>
<tr>
<th>Inoculation against Enteric Fever:</th>
<th>Correlation between Incubation and Inoculation.</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Hospital Staff...</td>
<td>...</td>
</tr>
<tr>
<td>II. Ladysmith Garrison...</td>
<td>...</td>
</tr>
<tr>
<td>III. Methven’s Column...</td>
<td>...</td>
</tr>
<tr>
<td>IV. Single Regiments...</td>
<td>...</td>
</tr>
<tr>
<td>V. Army in India...</td>
<td>...</td>
</tr>
<tr>
<td>Mean Value</td>
<td>...</td>
</tr>
</tbody>
</table>
Agricultural Research

• "Agricultural experiments on the same factor or group of factors are usually carried out at a number of places and repeated over a number of years. [...] The agricultural experimenter is thus frequently confronted with the results of a set of experiments on the same problem, and has the task of analysing and summarizing these." (Yates & Cochran, 1938)

• laid out methods that are still in use today (Cochran, 1937, 1943, 1954; Cochran & Carroll, 1953; Rao, Kaplan, & Cochran, 1981; Yates & Cochran, 1938)

• some core ideas:
  • estimates not equally precise (different variances)
  • compute weighted average of the estimates, with weights inversely proportional to the variances
  • estimates may be more variable than one would expect given their variances (heterogeneity)

Physics

• "Let us suppose that a given constant has been measured in several different ways, [...] it seems quite legitimate to combine by means of least squares the various results, i.e., to weight them according to their stated probable errors, and to derive the probable error in the final weighted average by the usual formulas." (Birge, 1932)

Glass (1976)

• based on his presidential address at the 1976 Annual Meeting of the AERA
• "Meta-analysis refers to the [...] statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings. 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."

Table 1

<table>
<thead>
<tr>
<th>SES Measure Consists of Indicators of</th>
<th>Average $t_{df}$</th>
<th>SES Measure Consists of Indicators of</th>
<th>Average $t_{df}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Income (only)</td>
<td>.315 (19)</td>
<td>Income &amp; Education</td>
<td>.230 (36)</td>
</tr>
<tr>
<td>Education (only)</td>
<td>-1.89 (116)</td>
<td>Income &amp; Occupation</td>
<td>3.32 (15)</td>
</tr>
<tr>
<td>Occupation (only)</td>
<td>.201 (65)</td>
<td>Education &amp; Occupation</td>
<td>.328 (30)</td>
</tr>
<tr>
<td>All Three</td>
<td></td>
<td>All Three</td>
<td>3.18 (27)</td>
</tr>
</tbody>
</table>

*Number of coefficients averaged in parentheses.
Early Meta-Analyses

- effects of interpersonal expectations on behavior (Rosenthal & Rubin, 1978)
- relation between class size and academic achievement (Glass & Smith, 1979)
- differential validity of employment tests for Black and White workers (Hunter et al., 1979)

Some Early Books (1985/1984)

But there were critics ...

- “mega-silliness” (Eysenck, 1978)
- “meta-analysis/shmeta-analysis” (Shapiro, 1994)
- “statistical alchemy for the 21st century” (Feinstein, 1995)

Systematic Reviews

- research synthesis as a scientific process
- based on replicable and systematic methods that are meant to “limit bias in the assembly, critical appraisal, and synthesis of all relevant studies on a specific topic” (Last, 2001)
- methods should be made explicit
- not necessarily supported by quantitative methods (can also use qualitative methods)

Cochrane and the EBM Movement

- Cochrane (1972). Effectiveness and efficiency: Random reflections on health services.
- advocated the use of RCTs to form the evidence base for clinical decision making

http://www.cochrane.org  http://www.campbellcollaboration.org

The Cochrane Library

Systmatic reviews of the effects of interventions in education, crime and justice, and social welfare to promote evidence-based decision-making.
Outcome Measures for Meta-Analysis
• commonly used outcome measures:
  • raw or standardized mean differences
  • risk differences, risk/odds ratios
  • correlation coefficients
  • means, proportions
  • reliability coefficients
  • ...

Observed vs. True Outcomes
• $y_i = \text{observed outcome in the } i\text{th study}$
• $\theta_i = \text{true outcome in the } i\text{th study}$
• $v_i = \text{sampling variance of } y_i$ (variability in estimates if one were to repeat the $i$th study under identical circumstances)
• assume: $y_i \sim N(\theta_i, v_i)$

Example: Standardized Mean Difference
• standardized mean difference:
  $$d = \frac{\bar{x}_1 - \bar{x}_2}{SD_p} \text{ is an estimate of } \theta = \frac{\mu_1 - \mu_2}{\sigma}$$
• sampling variance:
  $$v = \frac{1}{n_1} + \frac{1}{n_2} + \frac{d^2}{2(n_1 + n_2)}$$
• approximate 95% CI for $\theta$:
  $$d \pm 1.96\sqrt{v}$$

Pygmalion in the Classroom
• famous study by Rosenthal & Jacobson (1968)
• elementary school children were administered the “Harvard Test of Inflected Acquisition”
• randomly selected 20% of children were ‘identified’ as being intellectual ‘bloomers’
• ‘bloomers’ gained significantly more in total IQ (3.8 points) than control group children
• evidence how expectations can influence intellectual growth (self-fulfilling prophecy)
Pygmalion in the Classroom

Table 1

<table>
<thead>
<tr>
<th>Grade</th>
<th>Control</th>
<th>Experimental</th>
<th>IQ Points</th>
<th>Gain T1x &lt; 0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>48</td>
<td>70</td>
<td>19.34</td>
<td>0.02</td>
</tr>
<tr>
<td>2</td>
<td>49</td>
<td>69</td>
<td>16.00</td>
<td>0.05</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>20</td>
<td>12.59</td>
<td>1.84</td>
</tr>
<tr>
<td>4</td>
<td>65</td>
<td>62</td>
<td>11.06</td>
<td>0.03</td>
</tr>
<tr>
<td>Total</td>
<td>255</td>
<td>248</td>
<td>12.22</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Conflicting Results

• the study and its finding was readily accepted by some and harshly critiqued by others
• 14 years of additional research produced conflicting results
• Raudenbush (1984) conducted a meta-analysis of the existing evidence

Equal-Effects Model

- assume \( \theta_1 = \cdots = \theta_k = \theta \)
- then \( y_i \sim N(\theta, \tau^2) \)
- estimate \( \theta \) with:
  \[
  \hat{\theta} = \frac{\sum w_i y_i}{\sum w_i}
  \]
  where \( w_i = 1/v_i \)

Random-Effects Model

- but often true effects are not homogeneous (so-called Q-test can be used to test this)
- assume \( \theta_1 \sim N(\mu, \tau^2) \)
- then \( y_i \sim N(\mu, v_i + \tau^2) \)
- estimate \( \tau^2 \) and then \( \mu \) with:
  \[
  \hat{\mu} = \frac{\sum w_i y_i}{\sum w_i}
  \]
  where \( w_i = 1/(v_i + \hat{\tau}^2) \)

Results

- \( Q = 35.83, df = 18, p = .007 \)
- reject homogeneity assumption
- we find:
  \[
  \hat{\tau}^2 = 0.02 \\
  \hat{\mu} = 0.08
  \]
- 95% CI for \( \mu \):
  \( (-0.02, 0.18) \)
But that may not be the whole story …

- two phases to such experiments
  - first phase: induce expectation
  - second phase: test the expectancy hypothesis
- timing of first phase may be crucial
- if teachers had contact with children prior to expectancy induction, may not have an effect

Meta-Regression

- extension that allows inclusion of predictors ('moderators') in the models
- most general: mixed-effects meta-regression
- assume $\theta_i \sim N(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \tau^2)$
- estimate $\tau^2$ ('residual heterogeneity') and the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$

Some Misconceptions

- meta-analysis is objective
- need lots of studies for a meta-analysis
- meta-analysis is for synthesizing group differences or correlation coefficients
- a meta-analysis is a good first-year project for PhD students … well, maybe …
Some Recent Developments

- better methods for inference
- quantification of heterogeneity
- methods for specific types of data
- publication bias
- multilevel/multivariate models
- network meta-analysis

Publication Bias

- affects all review methods (not a problem specific to meta-analysis!)
- in fact, due to meta-analysis:
  - increased awareness of publication bias
  - development of systematic methods to detect and address publication bias
  - increased emphasis on the importance of trial registries and pre-registration

Multilevel Meta-Analytic Data

- multilevel structures can arise when we have multiple estimates for some higher clustering variable (paper, lab, research group, ...)

Multiple (Correlated) Outcomes

- multivariate data also arise when multiple outcomes are measured within the studies

Network Meta-Analysis

- mixed treatment comparisons meta-analysis
- esp. relevant for examining treatment effects
- often there are multiple treatments available for the same condition/disease
- studies comparing the effectiveness of these treatments form a network of comparisons

Star-Shaped Networks

Second-generation antiepileptic drugs in partial epilepsy

a: levetiracetam, b: gabapentin, c: lamotrigine, d: oxcarbazepine, e: tiagabine, f: topiramate, g: zonisamide, h: placebo
Goals of a Network Meta-Analysis

- synthesize evidence provided by all studies and treatment comparisons in one model
- obtain indirect evidence about comparisons that have not been examined head-to-head
- establish hierarchy of treatment effectiveness
- ...

Multilevel/Multivariate/Network MA

- analysis conducted with more complex mixed-effects models (e.g., Berkey et al., 1998; Konstantopoulos, 2011; Lu & Ader, 2004; Lumley, 2002; Salanti et al., 2008; Senn et al., 2013; van Houwelingen et al., 2002)
- Bayesian methods popular for network MA

References


References


Thank You!

Questions?