# Meta-Analysis Workshop

Michael LaValley
December 12th 2014
Villanova University

BOSTON UNIVERSITY

## Outline

- Hour 1
  - Definitions of systematic review and meta-analysis
  - Why do a meta-analysis?
  - Steps in performing a meta-analysis
  - Formulation of meta-analysis hypothesis
  - Systematic literature search
  - Effect size

## Outline

- Hour 2
  - Effect size
  - Fixed effect meta-analysis
  - Heterogeneity
  - Random effects meta-analysis

## Outline

- Hour 3
  - Meta-regression
  - Bias
  - Critiques and limitations of meta-analysis
  - Conclusions

## Definition of meta-analysis

- In empirical research, it is routine for more than one study to be done to address a single research question
- Standard analysis – take a single study without reference to other studies and attempt to reach a conclusion
- *Meta-analysis* – take a **set of studies** that address the same hypothesis and attempt to reach a conclusion

## Definition of meta-analysis

***Meta-analysis*** – the statistical analysis of a **collection** of studies

- Statistical (or quantitative) analysis rather than a narrative review
- Analysis is typically done at the study level
  - Analysis of study results
  - Study of studies

## Definition of meta-analysis

- Suppose that we are interested if alcohol consumption is associated with dating violence
  - We could conduct an observational (cross-sectional or case-control study) to estimate this association and test its statistical significance

  Or

  - We could collect a set of studies that have already evaluated this hypothesis and test the observed associations across the set of studies

## Definition of meta-analysis

- However, results usually vary among studies in a set
  - Positive or negative association?
  - Association statistically significant?
  - What is the magnitude of the association?
- How can we go though the set of studies and determine
  - The *average* association between alcohol consumption and dating violence
  - If the observed associations are consistent or inconsistent
  - If there are study factors associated with the observed association

## Definition of meta-analysis

- ***Systematic review*** is systematic assembly, appraisal, and synthesis of all relevant studies on a given topic
- ***Meta-analysis*** is often done as the statistical part of a *systematic review*
- The terms 'meta-analysis' and 'systematic review' are often used somewhat interchangeably
- But, there can be
  - Systematic reviews with no meta-analysis
  - Meta-analyses done outside of systematic reviews

## Definition of meta-analysis

- I will mainly focus on the statistical analyses used in meta-analysis
- But will also briefly cover other components of systematic reviews
  - Search for studies
  - Extraction of study results
  - *An analysis is only as good as the data on which it is based*

## Definition of meta-analysis

- Typically meta-analysis is done on the **summary results** of the assembled collection of studies
  - Data used in meta-analysis are the study results not the individual subject data
  - Exception is individual patient data (IPD) meta-analyses

## Definition of meta-analysis

- Usual goal for meta-analysis
  - Analyze results from a collection of primary studies to arrive at a *combined* or *average* estimate of effect across the studies
- Another goal (sometimes)
  - Look for factors to explain differences between study results
  - Usually done with a stratified or regression analysis

From a meta-analysis on the effect of initial mean severity of depression and the effect of antidepressant medications in randomized controlled trials that had been submitted to the US FDA.

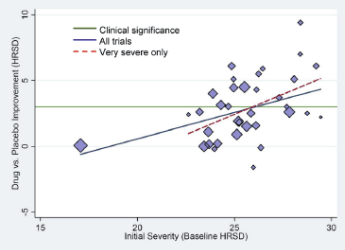



**Figure 4.** Mean Drug–Placebo Difference Scores as a Function of Initial Severity

Plotted values are sized according to their sample sizes (*n*); the green line represents the NICE clinical significance criterion. The solid blue regression line represents the trend across all 35 trials; the dashed red line represents the trend excluding the left-most observation.
doi:10.1371/journal.pmed.0050045.g004

## Definition of meta-analysis

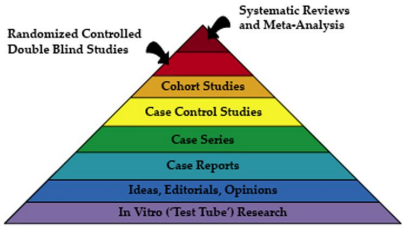



**Figure 1.** Evidence pyramid graphic, from SUNY down-state medical research library of Brooklyn, EBM tutorial

## Definition of meta-analysis

- Typical view of the hierarchy of evidence
- I find this view to be a little to simple
  - There is no privileged road to the truth
- Every type of study has its own pros and cons
  - The best study type is the study type that best addresses your question


Figure 1. Evidence pyramid graphic, from SUNY down-state medical research library of Brooklyn, EBM tutorial

## Definition of meta-analysis

- Meta-analysis is a retrospective, observational analysis
  - Included studies may or may not be prospective, or experimental
- But a meta-analysis
  - Uses previously conducted research – retrospective
  - Does not assign interventions – observational

## Definition of meta-analysis

- Meta-analysis is subject to the same concerns that apply to other retrospective, observational study types
  - Selection bias
    - Are the included studies representative of the research in an area?
  - Confounding
    - Are there other factors associated with both the factor of interest and the outcome?

## Why do a meta-analysis?

- To provide a summary of the research findings regarding a hypothesis
  - The sheer numbers of research studies require quantitative methods to summarize research findings
    - Chan and Altman found 519 randomized clinical trials indexed in PubMed for the month of December 2000
    - Bastian, Glasziou and Chalmers found 75 trials (and 11 systematic reviews) published each **day** in 2010

## Why do a meta-analysis?

- To increase significance of an effect
  - Individual studies may have low power to reach statistical significance for a small but meaningful effect
  - By using many studies the sample size can be increased to where small effects reach statistical significance

## Reasons for meta-analysis

- To *explore* possible reasons for conflicting results among primary studies in a research area
  - May generate new hypotheses to be tested in further primary studies
  - There isn't a way to do this kind of analysis across separate studies outside of meta-analysis

Epidemiologic Reviews

DOI: 10.1093/epirev/mxr027

**Does the Alcohol Make Them Do It? Dating Violence Perpetration and Drinking Among Youth**

**Emily F. Rothman*, Luz McNaughton Reyes, Renee M. Johnson, and Michael LaValley**

* Correspondence to Dr. Emily F. Rothman, Department of Community Health Sciences, Boston University School of Public Health, 801 Massachusetts Avenue, Floor 3, Crosstown Center, Boston, MA 02118 (email: erothman@bu.edu).

*Accepted for publication September 13, 2011.*

Strong evidence links alcohol use to partner violence perpetration among adults, but the relation between youth alcohol use and dating violence perpetration (DVP) is not as well studied. The authors used meta-analytic procedures to evaluate current knowledge on the association between alcohol use and DVP among youth. The authors reviewed 28 studies published in 1985–2010; most (82%) were cross-sectional. Alcohol use was measured in 3 main ways: 1) frequency or quantity of use, 2) frequency of heavy episodic drinking, or 3) problem use. Collectively, results support the conclusion that higher levels of alcohol use are positively associated with youth DVP. With fixed-effects

## Forest plots

- Forest plots provide a way to display most of the results of a meta-analysis in a single figure
- This plot has become a standard display in meta-analysis

## Forest Plot

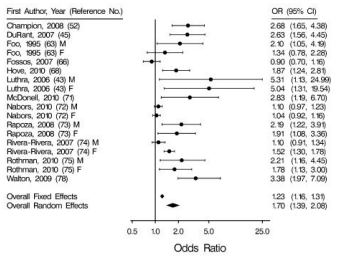
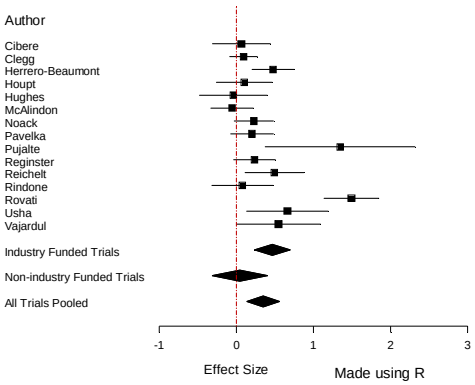
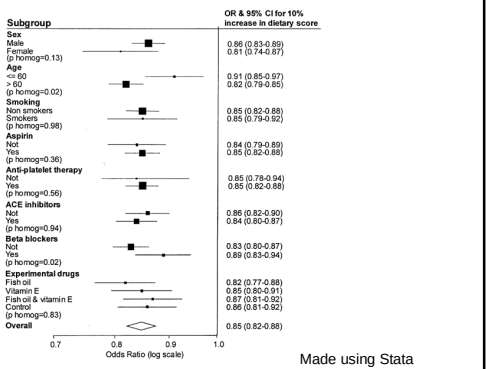



**Figure 2.** Forest plot illustrating relative strength of associations between frequency/quantity of alcohol use and dating violence. CI, confidence interval; F, results for females; M, results for males; OR, odds ratio.

Made using SAS




Made using R




Made using Stata

## Forest plots

- Structure of a forest plot
  - Primary studies
    - *Effect size* for a primary study shown with a symbol (square, rectangle, or circle)
    - Confidence intervals are shown by whiskers extending from the effect size
  - The combined estimate
    - Often shown by a diamond
      - Highest point shows combined estimate
      - Width shows confidence interval width

## Overview of Doing Meta-analysis

1. Acquire study results from a set of studies
2. Combine study results across set to obtain summary effect and confidence interval
3. Look for problems!
   - Heterogeneity
   - Bias
4. Regression or stratification for influence of study level factors?

## Formulating a Hypothesis

- An acronym that is helpful in formulating a targeted hypothesis is ***PICOS***
  - Population (or Participants)
  - Interventions
  - Comparator
  - Outcomes
  - Study Designs

## Systematic searches

- A crucial part of a systematic review is a thorough search for relevant studies
- When meta-analyses reach differing conclusions, it is often due differences in which studies were included
- The methods used for the search should be explicitly described
  - Listed in the meta-analysis protocol
  - Given in the methods section of a paper or report
  - Put in an online appendix

## Systematic searches

- Places to search for Published Studies
  - Electronic citation databases
    - PubMed, Embase, CINAHL, PsychLIT, Cochrane Collaboration Central Register of Controlled Clinical Trials, Web of Science
  - Web search
    - Google scholar
  - Hand search of particularly relevant journals?
  - Citation lists of previously obtained articles and reviews – snowball search

## Systematic Searching

- Step 1 – Scoping Search
- Step 2 – Develop a formal list of search terms
  - Start with PICOS
  - Iterate and improve
- Step 3 – Snowball
  - Evaluate reference lists of found primary studies and reviews
  - Use citation indexes to look for papers citing the primary studies and previous reviews

## Systematic Searching

- Scoping Search
  - A preliminary search to get a sense of what is out there
    - Any systematic reviews?
    - Rough numbers of hits to evaluate size of the literature
  - Should give a sense of whether the proposed meta-analysis is worthwhile
    - Has this meta-analysis already been done?
    - Are there gaps in the literature?
    - Will there be enough studies for a meta-analysis?

## Systematic Searching

- Scoping Search
  - Develop the initial research question
  - Restrict to one database (most likely PubMed)
  - Use just a few key terms
  - Think of alternative terms (synonyms)
  - Search for terms separately and combine with Boolean operators (AND/OR/NOT)

## Systematic Searches

- In doing the formal search it is important to balance the precision with the recall
  - *Precision* of a search = proportion of citations retrieved that are relevant to the meta-analysis
  - *Recall* of a search = proportion of the potentially relevant citations that are retrieved

## Systematic Searches

- Suppose that there are 50 studies that are relevant for your meta-analysis
  - A search that returns 10,000 citations may have great recall (include all 50 relevant studies – 100%) but poor precision (0.5% of citations are relevant)
  - A search that returns 2 citations may have great precision (100% of studies are relevant) but poor recall (only found 4% of relevant studies)

## Systematic Searches

- In published meta-analyses, we only see the final search strategy, but it should be built up iteratively
- If the scoping search identified some studies that you would want to include
  - Did the search identify all these studies? (recall)
  - Did the search bring in too many irrelevant citations? (precision)

## Systematic Searches

- Based on the initial search results, may need to add more search terms to boost either the recall or precision
  - Add more synonyms for key terms?
  - Use more 'AND' operators to reduce numbers of citations?

## Systematic Searches

- It is worthwhile to discuss the search with a reference librarian
  - They will have a lot more domain knowledge about the databases and the mechanics of the searches
    - Use of wildcard or truncation operators (smoke* or smoke$ instead of smoking)
    - Use of MESH terms in MedLine
    - Subheadings
    - 'Exploding' terms

## Systematic searches

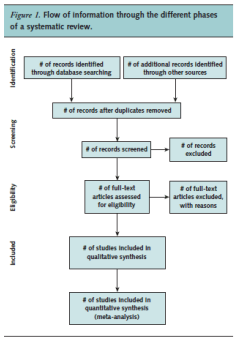
- Usually a lot of citations are retrieved from database searches. These are weeded down in stages
  1. Looking at the titles
  2. Looking at the abstracts
  3. Looking at the paper for those citations where the abstract seems promising

## Inclusion/Exclusions

- In weeding down the studies, the primary tool are the inclusion and exclusion criteria
  - Only Subjects drawn from a certain population?
  - Only certain forms of the outcome to be included?
    - Particular scales
  - Only accept certain forms of treatment or exposure?
- Want criteria to be clear-cut

## Snowball searches

- A ***snowball*** search can be done after the database search
  - Look for citations from the primary studies and any relevant systematic reviews
  - Look for papers citing these primary studies and systematic reviews
  - This can be done easily with the web of science database
- Snowball searches can be especially helpful if the formal database searches have produced limited results




Flow chart for reporting searches from PRISMA statement

http://www.prisma-statement.org/

## Extraction

- Once the primary studies have been identified and the text is available, the summary results need to be *extracted* from the study reports
- Also called *coding*
- Extracting study results is often much more difficult than one would expect

## Extraction

- Questions that arise in extraction of primary study results
  - Which of multiple measures of the outcome should be used for a study?
  - What time interval for study results?
  - Use change from baseline or end of study results?
  - What if there are more than two groups?
- Unfortunately, these details can matter!

## Extraction

- To obtain efficacy data from the study reports located by searching the literature, a **standardized form** should be used
  - Reduce the time needed to extract the data
  - Ensures extraction is done in a uniform manner
  - List the page, table and figure numbers where values are taken from for later checking

## Extraction

- In general, it is best if more than one person extracts all the study results
  - Allows comparisons of data between extractors
  - Differences can be adjudicated
  - Should ensure a more reliable extraction and better data
- Trade off between efficiency and data reliability

## Extraction of Continuous Outcomes

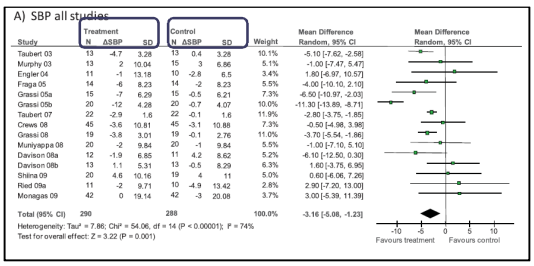
- Need 6 numbers for extraction of continuous outcomes

| Group | Sample Size | Mean | Standard Deviation |
|---|---|---|---|
| Experimental | $n_e$ | $\bar{X}_e$ | $s_e$ |
| Control | $n_c$ | $\bar{X}_c$ | $s_c$ |

- The standard deviation is a **key** piece of information for creating the effect size measures
  - Often not reported

## Extraction of Continuous Outcomes

- If a different measure of variability is reported then can often convert it into the standard deviation
- If no variability estimate is presented at all
  - Contact the authors
  - Generally shown to be better to impute SD than to not use the study at all
  - Usually use SDs from the most similar studies
    - Same time frame
    - Same outcome scale
    - Same population



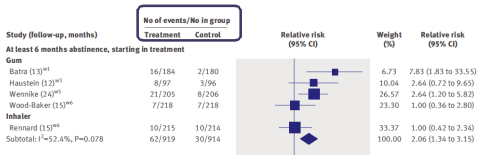

Continuous data extracted in meta-analysis on effect of chocolate on systolic blood pressure

## Extraction of Binary Outcomes

- Need 4 counts for extraction of binary (*proportion*) outcomes

| Group | Event Count | No Event Count | Total Count | Event Percentage |
|---|---|---|---|---|
| Experimental | a | b | $n_e$=a+b | $p_e$=a/$n_e$ |
| Control | c | d | $n_c$=c+d | $p_c$=c/$n_c$ |

- Need a, b, c, d or the basic two-by-two table
- Could get a, c, $n_e$, $n_c$
- Or $n_e$, $n_c$, $p_e$, $p_c$




Binary data extracted in meta-analysis on proportion of subjects on nicotine replacement therapy who are abstinent for 6 months compared to controls

## Effect Sizes

- The first statistical step in meta-analysis is to take the extracted data *from each study* and calculate
  - **Effect size**
  - Associated standard error

## Effect Sizes

- Effect size is a measure of the distance between the null hypothesis and observed data
  - Effect size measures the degree to which the observed data are inconsistent with the null hypothesis

## Effect Sizes

- Effect size is often used to provide a measure of the **clinical significance** of a study result
  - Statistical significance without clinical significance may denote an effect that is not worth pursuing
  - Clinical significance without statistical significance is an interesting result that might be only a chance finding

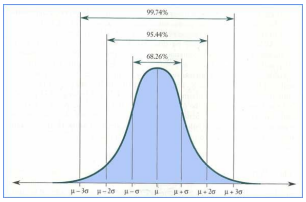
## Effect Sizes

- Effect size measures are based on type of outcome data
  - Continuous
  - Discrete
  - Time to event
- The most *common* effect measures listed in the following slides
- But, almost any statistic with a confidence interval can be employed as effect sizes

## Normality of Effect Sizes

- For inverse variance meta-analysis methods it will be useful for the effect sizes to have a roughly normal distribution

Normal Distribution




## Normality of Effect Sizes

- To obtain *approximate* normality for ratio measures we often need the natural log transformation for the measure
  - Odds ratio
  - Relative risk
- Other transformations may be useful to normalize other effect measures
  - Fisher's z-transformation for correlations
  - Arcsine or Freeman-Tukey transformation for percentages

| Type of Outcome | Effect Size |
|---|---|
| Continuous | (Fisher's z-transformation) Correlation |
| | Mean Difference (MD) |
| | Standardized Mean Difference (SMD) |
| Binary | (Log) Odds Ratio (OR) |
| | (Log) Relative Risk (RR) |
| | Risk Difference (RD) |
| Time to Event | (Log) Rate Ratio<br>(Log) Hazard Ratio |

## Continuous Effect Size

- Fisher's z-transformation of the correlation
  - $z = \frac{1}{2} ln \left(\frac{1+r}{1-r}\right)$
- Where $r$ is the observed correlation from $n$ observations
- $z$ has a standard error of
  - $SE_z = \frac{1}{\sqrt{n-3}}$

## Continuous Effect Sizes Mean Difference (MD)

| Group | Sample Size | Mean | Standard Deviation |
|---|---|---|---|
| Experimental | $n_e$ | $\overline{X}_e$ | $s_e$ |
| Control | $n_c$ | $\overline{X}_c$ | $s_c$ |

Estimator: $\mathbf{MD} = \overline{\mathbf{x}}_e - \overline{\mathbf{x}}_c$

Standard Error: $\mathbf{se}_{MD} = \sqrt{\frac{s_e^2}{n_e} + \frac{s_c^2}{n_c}}$

95% Confidence Interval: $MD \pm 1.96 * se_{MD}$

## Continuous Effect Sizes
## Standardized Mean Difference

| Group | Sample Size | Mean | Standard Deviation |
|---|---|---|---|
| Experimental | $n_e$ | $\bar{X}_e$ | $s_e$ |
| Control | $n_c$ | $\bar{X}_c$ | $s_c$ |

Estimator

$$SMD = f\,\frac{\bar{X}_e - \bar{X}_c}{s}$$

where:

$$s = \sqrt{\frac{(n_e-1)s_e^2 + (n_c-1)s_c^2}{n_e + n_c - 2}}$$

$$f = \frac{4(n_e + n_c - 2) - 4}{4(n_e + n_c - 2) - 1}$$

Standard Error and 95% Confidence Interval

$$se_{SMD} = \left[\frac{n_e + n_c}{n_e n_c} + \frac{d^2}{2(n_e + n_c)}\right]^{1/2}$$

$$SMD \pm 1.96 * se_{SMD}$$

## Continuous Effect Sizes

- When to use mean difference (MD)
  - When primary studies have comparable outcome measures
  - Gives results with familiar units
- When to use standardized mean difference (SMD)
  - When primary studies use different measures of the same construct (different scales)
  - Gives results in number of standard deviations – unit free

## Continuous Effect Sizes

- Normand meta-analyzed 9 studies comparing length of stay following a stroke in specialty stroke units to standard rehabilitation units
  - Combined MD was -15.0 days for specialist care, with 95% confidence interval from -31.4 to 1.4 days
  - Combined SMD was -0.56 standard deviations for specialist care, with 95% confidence interval from -1.15 to 0.02 standard deviations

## Discrete Effect Size
## Log Odds Ratio ($L_o$)

| Group | Event Count | No Event Count | Total Count | Event Percentage |
|---|---|---|---|---|
| Experimental | a | b | $n_e$=a+b | $p_e$=a/$n_e$ |
| Control | c | d | $n_c$=c+d | $p_c$=c/$n_c$ |

Estimator: $o = \dfrac{p_e/(1-p_e)}{p_c/(1-p_c)}$ $\quad L_o = \ln(o)$

Standard Error: $s_{L_o} = \left[\dfrac{1}{n_e p_e(1-p_e)} + \dfrac{1}{n_c p_c(1-p_c)}\right]^{1/2}$

100(1- $\alpha$)% CI: $L_o \pm Z_{\alpha/2}\, s_{L_o}$ $\quad \exp(L_o \pm Z_{\alpha/2}\, s_{L_o})$

## Discrete Effect Size
## Log Relative Risk ($L_r$)

| Group | Event Count | No Event Count | Total Count | Event Percentage |
|---|---|---|---|---|
| Experimental | a | b | $n_e$=a+b | $p_e$=a/$n_e$ |
| Control | c | d | $n_c$=c+d | $p_c$=c/$n_c$ |

Estimator: $r = p_e/p_c$ $\quad L_r = \ln(r)$

Standard Error: $s_{L_r} = \left[\dfrac{1-p_e}{n_e p_e} + \dfrac{1-p_c}{n_c p_c}\right]^{1/2}$

100(1- $\alpha$)% CI: $L_r \pm Z_{\alpha/2}\, s_{L_r}$ $\quad \exp(L_r \pm Z_{\alpha/2}\, s_{L_r})$

## Discrete Effect Size
## Risk Difference (RD)

| Group | Event Count | No Event Count | Total Count | Event Percentage |
|---|---|---|---|---|
| Experimental | a | b | $n_e$=a+b | $p_e$=a/$n_e$ |
| Control | c | d | $n_c$=c+d | $p_c$=c/$n_c$ |

Estimator: $RD = p_E - p_C.$

Standard Error: $SE_{RD} = \sqrt{\dfrac{p_E*(1-p_E)}{n_E} + \dfrac{p_C*(1-p_C)}{n_C}}.$

100(1- $\alpha$)% CI: $RD \pm Z_{\alpha/2}\, SE_{RD}$

## Discrete Effect Sizes

- Risk difference not used too often
  - Empirically shows more variability than the OR or RR
  - Additive effect instead of multiplicative

## Discrete Effect Sizes

- Odds ratio has been the favored effect measure in the past
  - Odds ratio is more symmetric if outcome event is changed
- Relative risk given more emphasis in evidence-based-medicine
  - Relative risk easier to interpret

## Converting Effect Sizes

- Sometimes a set of studies will use a mixture of outcome types
  - For example
    - Pain measured on continuous scales for most studies
    - Pain measured as a dichotomous response in a few studies
- There are conversion formulas to convert from one type of effect size to another
  - *compute.es* package in R is great for this!

## Basics of Combining Results

- In meta-analysis we **combine** results across multiple studies while preserving the studies as distinct
- In some papers *pooling* is used to refer to putting all the data together as if from one big study
  - This does not preserve the studies as distinct unless analysis is stratified by study
- But in some papers meta-analysis results are described as pooled – **be careful about terminology**!

## Basics of Combining Results

- In the meta-analysis approaches that we consider, you can think of the study as a ***matching*** or stratification factor
  - If the researchers have done a good job exposed and control subjects will be well-matched within each study
  - But exposed subjects and control subjects can be different from one study to the next
- Generally we try to avoid indirect comparisons between subjects in one study with subjects in a different study

## Basics of Combining Results

- Key assumption – **independence** of study results
  - There should be no subjects who are included in more than 1 result in the set to be combined
- One study can contribute more than 1 result if analyses presented for non-overlapping subgroups
  - Men and women
  - Obese and normal weight subjects

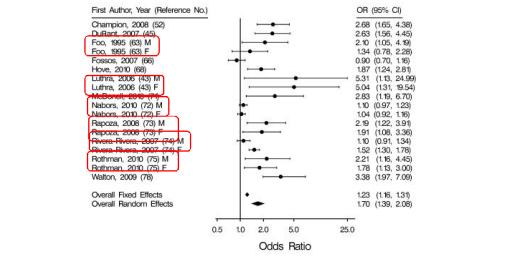



**Figure 2.** Forest plot illustrating relative strength of associations between frequency/quantity of alcohol use and dating violence. CI, confidence interval; F, results for females; M, results for males; OR, odds ratio.

## Combining Study Results

- Combine effect sizes from studies
  - Sampling error gives ***within study variability*** (due to sample size)
  - Is there also ***variability between*** studies?
    - If not, ***fixed effect*** analysis
    - If so, ***random effects*** analysis
  - Both analyses compute weighted averages of the study effects, but the weights are different

## Fixed Effect Combining

- Fixed effect meta-analysis ***assumes*** that all studies are estimating the same underlying parameter $\delta$
  - $\delta$ is the true size of effect in the population
    - The expected value for each of the effect sizes in the set of primary studies is $\delta$
  - The **only** reason studies have different effect estimates is random sampling error
    - If a primary study had infinite sample size, then the effect size estimate would exactly equal $\delta$

## Inverse Variance Weighting

- $d_i$ is assumed to be **normally distributed** around $\delta$ with variance due to within study variability

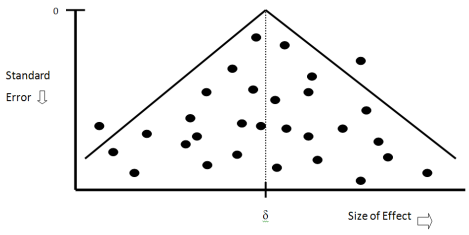
$$d_i \sim N(\delta, v_i)$$

- The ***within study variability*** $v_i$ is the square of the standard error of the estimate $d_i$

## Funnel Plot

Since almost all of the confidence intervals need to contain $\delta$, study effect sizes need to follow the funnel shape shown below when plotted against the standard error

As standard error decreases, effect estimates converge on the true effect




$\delta$ represents the true size of the effect

## Inverse Variance Weighting

- Since all studies estimate $\delta$ it makes sense to **average** the study estimates for the combined estimate

$$\bar{d} = w_1 d_1 + w_2 d_2 + \cdots + w_k d_k$$

- $w_i$ is the **weight** given to study i
  - $0 \leq w_i$ for each i
  - $w_1 + \ldots + w_k = 1$

## Inverse Variance Weighting

- Different choices of $w_i$ are possible
  - Examples: $w_i=1/k$, $w_i=n_i/(n_1+\ldots+n_k)$
- The choice of $w_i$ affects the standard error of the combined estimate of $\delta$
- The **best** choice of $w_i$ will minimize the standard error of the weighted average

## Inverse Variance Weighting

- The best choice – weights should be proportional to the study precision

$$w_i \propto \frac{1}{SE(d_i)^2}$$

- More precise studies (more subjects, smaller standard error) get more weight
- Less precise studies (fewer subjects, larger standard error) get less weight

## Inverse Variance Weighting

- Here the *variance* is the within-study variance of the effect estimate: $Var(d_i) = SE(d_i)^2$
- The best weight to use is

$$w_i = \frac{\frac{1}{SE(d_i)^2}}{\frac{1}{SE(d_1)^2} + \frac{1}{SE(d_2)^2} + \cdots + \frac{1}{SE(d_k)^2}}$$

  - The bottom of the fraction is the same for every study – ensures that the weights sum to 1

## Inverse Variance Weighting

- The standard error of the combined estimate

$$SE_{\bar{d}} = \sqrt{\frac{1}{\sum_{i=1}^{k} \frac{1}{SE(d_i)^2}}}.$$

## Other Fixed Effect Methods

- The inverse variance weighting method can be used to combine any effect estimates that are roughly normally distributed
  - Usually requires the natural log transformation for binary data
- There are other fixed effect methods developed specifically for **binary** outcomes that do not require rough normality of the effect estimates
  - Cochran-Mantel-Haenszel method
  - Peto method

## Cochran-Mantel-Haenzel

- Familiar method used to combine results across k 2-by-2 contingency tables
  - Usually the tables are for k strata within a single study
    - k centers in a multi-center trial
    - k risk groups in a cohort study
  - But, this method can also be used for k different primary studies
- In this method averaging is done directly on the non-logged odds ratio (or relative risk) with different weights than inverse variance

## Peto Method

- The Peto method is another meta-analysis method for binary data and has been used extensively
- Peto method is based on an approximation to the likelihood
  - Expressed as a difference between observed and expected counts
  - Estimates the combined log odds ratio

## Zero Counts

- Inverse variance weighting approach with 0 counts in a primary study table
  - With the **log odds ratio** or **log relative risk** there are problems with division by 0 or having a log(0)
    - Often ½ is added to all cells in each table if this is a problem
- Inverse variance methods not considered to be the best choice for rare outcomes

## Zero Counts

- The Cochran-Mantel-Haenszel method can deal with one cell with a zero count, but not a row or column
- The Peto method is fine with tables with 0 cells
  - Become the standard fixed effect method for sparse tables
  - But also implicitly excludes tables with a row or column of 0 cells

Higgins et al., JRSS 2009

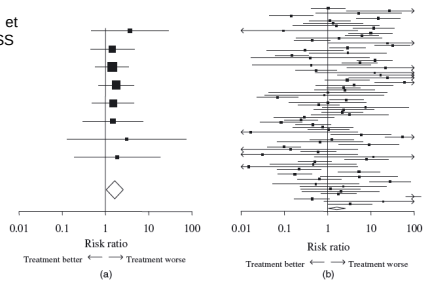



**Fig. 1.** Estimates with 95% confidence intervals for (a) a genuine meta-analysis and (b) an artificially constructed meta-analysis with identical results for the mean of a random-effects distribution

The combined estimate and its 95% confidence interval may not be sufficient to characterize a set of studies

## Homogeneity

- Assumption of **homogeneity**
  - This key assumption for the fixed effect meta-analysis approach is that all the primary studies are estimating **the same** population quantity
  - Every primary study provides an effect size estimate $d_i$ of $\delta$ - a single population effect size

## Homogeneity

- Under homogeneity, it makes sense to average effect size estimates from all the studies to obtain a precise combined estimate of $\delta$
- But, homogeneity is just an assumption, and it can be incorrect for a particular set of studies

## Heterogeneity

- If homogeneity does not hold, then the data are ***heterogeneous***
  - Some of the primary studies are estimating **different** population quantities

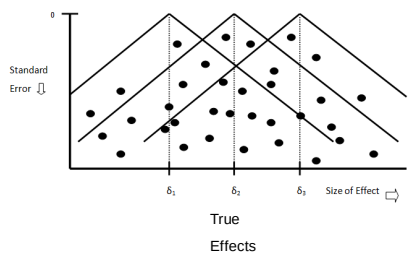
## Heterogeneity

- With heterogeneity there is no single population effect size $\delta$
  - Each study effect size $d_i$ estimates a **study-specific** population effect $\delta_i$
  - The effect size for study i approaches $\delta_i$ as the sample size increases
- Some of the $\delta_i$ might be the same, but not all of them

## Heterogeneity

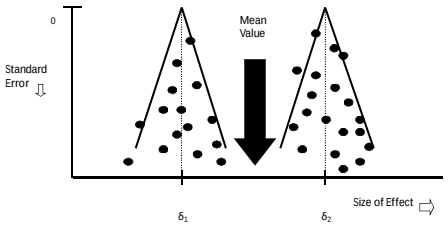
With heterogeneity there could be several funnels in the plot of study effect by standard error.

As standard error decreases, effect estimates do not need to converge.



## Heterogeneity

Excessive between study variability may make combined estimate meaningless if too much variability between study effects




Effect of extreme heterogeneity – the mean doesn't characterize either group or any observation

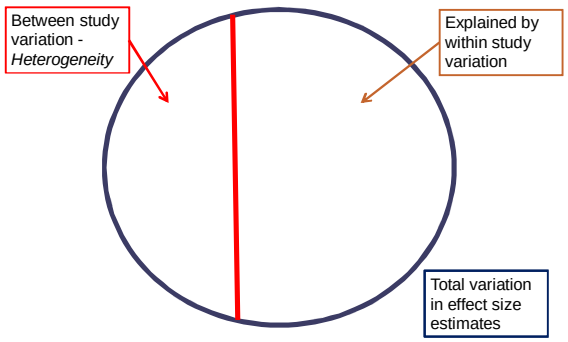
## Heterogeneity

- With heterogeneity,
  - Not all of the studies estimate exactly the same thing
  - It is not clear that combining the study effect sizes produces a meaningful estimate of any population value
- Need to consider if it is a good idea to combine study estimates
  - Combined estimate might not reflect any study in the population

## Heterogeneity

- Because statistical tests to detect heterogeneity usually have low power, do not conclude that a null result rules out any possibility of heterogeneity
- To supplement the tests we need descriptive, non-test based, diagnostic procedures for heterogeneity

## Partition the Total Variation




## Testing for Heterogeneity

- The main test is called Cochran's Q-test or just the Q-test
- Differences between the study effect size and the fixed effect combined effect size are squared (*squared deviations*)
- The squared deviations are weighted by the inverse of the within study variances and summed over all the studies

## Testing for Heterogeneity

- The formula for the Q-test statistic is

$$Q = \frac{1}{SE_{d1}^2} * \left(d_1 - \bar{d}\right)^2 + \frac{1}{SE_{d2}^2} * \left(d_2 - \bar{d}\right)^2 + \cdots + \frac{1}{SE_{dk}^2} * \left(d_k - \bar{d}\right)^2$$

- $\bar{d}$ is the fixed effect combined effect size
- $d_i$ is the effect size estimate for study i
- $SE_{di}$ is the standard error of $d_i$

## Testing for Heterogeneity

- The Q-statistic provides a test for the presence of **between-study variation**
  - $H_0$: homogeneity – no between study variation
  - $H_A$: heterogeneity – between study variation

## Testing for Heterogeneity

- With homogeneous studies
- Q has an approximate chi-square distribution with k-1 degrees of freedom (df)
- Q is compared to the upper 0.05 cutoff for a chi-square with k-1 df to determine if the heterogeneity is significant

## Estimating Heterogeneity

- The Q-test is also the basis of some descriptive measures of heterogeneity
  - H and $I^2$
  - Diagnostic plot

## H Statistic

- H is the square root of the Q-statistic divided by its degrees of freedom

$$H = \sqrt{\frac{Q}{k-1}}$$

- Under homogeneity, the mean of H is 1
- If H < 1, it is set equal to 1

$$H = \max\left(\sqrt{\frac{Q}{k-1}}, 1\right)$$

## H Statistic

- Higgins and Thompson suggest
  - If H>1.5 there should be considerable caution about the effect of heterogeneity on the meta-analysis
  - If H≤1.2 there should be little concern
- With only a few studies, the Q-test might not be significant while H is large (above 1.5)
- With many studies the Q-test might reach significance while H is small (below 1.2)

## $I^2$ Statistic

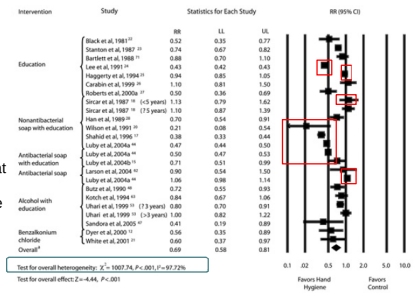
- $I^2$ is a transformation of H that estimates the proportion of the total variability in effect size that is due to between study variation

$$I^2 = \frac{H^2 - 1}{H^2}$$

## $I^2$ Statistic

- $0 \leq I^2 \leq 1$ and can be interpreted as a proportion (similar to $R^2$ in linear regression)
  - Values of $I^2$ close to 0 suggest that there is little between study variation (heterogeneity) in the data
  - Values of $I^2$ close to 1 indicate that between-study variation dominates within study variation and combined estimates may not be useful

Notice the studies with 95% confidence intervals that don't intersect the combined value




## $I^2$ Statistic

- $I^2$ has become a preferred measure of heterogeneity
- Higgins et al. recommend the following guidelines for interpreting $I^2$
  - **Low** heterogeneity: $I^2 \leq 0.25$
  - **Moderate** heterogeneity: $0.25 < I^2 < 0.75$
  - **High** heterogeneity: $0.75 \leq I^2$

## $I^2$ Statistic

- Keep in mind that the $I^2$ value is an ***estimate***
  - Not a very precise estimate unless the number of primary studies is quite large
  - Even if $I^2$ is estimated to be 0, it may not rule out substantial heterogeneity
- Also, $I^2$ is the proportion of variation in effect sizes that can't be explained by within study variation
  - If the within study variation is large (studies have low precision) then $I^2$ will be low even if the study results are quite variable

## Analysis in the Presence of Heterogeneity

- We expect some level of heterogeneity, so it makes sense to develop an analysis to account for moderate heterogeneity
- A model that incorporates between study variability was first proposed by DerSimonian and Laird and this is the basis of ***random effects*** meta-analysis
- The combined effect is still a **weighted average**, but the weights are not the fixed effects weights

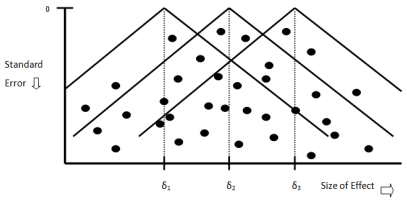
## Random Effects Model

- In the random effects model each study is considered to estimate a slightly different population effect size $\delta_i$
- I think of the $\delta_i$ as being different due to design differences among the studies
- The $\delta_i$ are called **random effects**

## Random Effects Model

With heterogeneity there could be several funnels in the plot of study effect by standard error.

As standard error decreases, effect estimates do not need to converge.




True Effects

## Random Effects Model

- To make a workable model, the random effects are **assumed** to follow a normal distribution with
  - Mean = $\delta$
  - Variance = $\tau^2$
- $\tau^2$represents the **between-study variability** or **heterogeneity**

## Random Effect Model

- This is a **2-level** model
  - **Level 1**: study effect size $d_i$provides an estimate of $\delta_i$the population effect size using the design set-up of study i

$$d_i \sim N(\delta_i, v_i)$$

  - $v_i = s_i^2$ is the **within study variance**

## Random Effect Model

- **Level 2**: population effect sizes $\delta_i$ have a normal distribution around the central value $\delta$ with variance $\tau^2$

$$\delta_i \sim N(\delta, \tau^2)$$

- $\tau^2$ is the **between-study** variance

## Random Effect Model

- In this 2-level model
  - The $d_i$ study effect sizes follow a normal distribution with
    - $mean = \delta$
    - $variance = s_i^2 + \tau^2$
  - The $d_i$ distribution is still normal, but with increased variance
    - This is why a normal distribution was used for level 2

## Random Effects Model

- The variability of an effect size estimate $d_i$ has two components in this model
  - Variability of $d_i$ as an estimator of $\delta_i$ – the ***within*** study variability $v_i = s_i^2$
  - Variability of $\delta_i$ around $\delta$ – the ***between*** study variability $\tau^2$

## Random Effects Model

- This is a type of Bayesian model as there is a distribution placed on some of the unknown parameters – the $\delta_i$ random effects
  - The parameters $\delta$ and $\tau^2$ are called hyper-parameters and they specify the distribution of the $\delta_i$ parameters
- Since we use the $s_i^2$ as if they were known values and not just estimates this is called an **Empirical Bayes** model

## Random Effects Model

- If there is no between study variability, then $\tau^2 = 0$
  - Variance of the random effects combined estimate will be the same as that of the fixed effects estimate
- Later, we will consider methods of obtaining estimates of $\tau^2$
- For now, just assume that we know $\tau^2$

## Random Effects Model

- Want the study weights reflect both within and between study variation
- For study i, the random effects weight is proportional to the reciprocal of the **sum** of the study standard error squared and the between study variability

$$w_i^* \propto \frac{1}{\left(s_i^2 + \tau^2\right)}$$

## Random Effects Model

- To get these weights to add to 1

$$w_i^* = \frac{\frac{1}{s_i^2 + \tau^2}}{\frac{1}{s_1^2 + \tau^2} + \frac{1}{s_2^2 + \tau^2} + \cdots + \frac{1}{s_k^2 + \tau^2}}$$

- I use $w_i^*$ to denote the random effects weights to keep them separate from the fixed effects weights

## Random Effects Model

- So, in a random effects model

$$\overline{d^*} = w_1^* d_i + w_2^* d_2 + \cdots + w_k^* d_k$$

- And the standard error of the random effects combined estimate is

$$SE_{\overline{d^*}} = \sqrt{\frac{1}{\sum \frac{1}{s_i^2 + \tau^2}}}$$

## Random Effects Model

- Random effects weights (w*) tend to give a more equal weighting to the studies in a meta-analysis than do the fixed effect weights
  - Most precise studies are down-weighted in RE meta-analysis compared to FE
  - Least precise studies are given more weight in RE meta-analysis compared to FE

## Random Effects Model

- In the extreme case, if $\tau^2$ becomes much larger than all of the within study variances ($s_i^2$), the random effects weights approach equal weighting for the studies (1/k)
- The random effects weighting for studies lies somewhere between the fixed effects inverse-variance weighting and equal weighting

## How to Estimate $\tau^2$

- The typical approach to estimate $\tau^2$ is the **DerSimonian and Laird** method of moments estimate
  - Based on the Q-test statistic
  - Implemented in most meta-analysis software

$$\tau^2 = \frac{Q - (k-1)}{\left(\sum \frac{1}{s_i^2}\right) - \left(\left(\sum \frac{1}{s_i^4}\right) \Big/ \left(\sum \frac{1}{s_i^2}\right)\right)}$$

## How to Estimate $\tau^2$

- Problem with the method of moments approach
  - The estimate of $\tau^2$ depends on the fixed effects estimate of $\delta$
  - But, the fixed effects estimate of $\delta$ is based on there being **no** between study variability
- Also, if Q < k-1, then the DL estimator can be negative

## How to Estimate $\tau^2$

- An alternative approach is the Sidik and Jonkman (SJ) estimator
- SJ avoid the possibility of negative values by looking at ratios of between to within study variance
- The SJ approach has worked better than DL in most simulation studies
- But DL remains the default in most meta-analysis packages

## Random Effects with Binary Data

- For the inverse variance random effect model, the first level required that $d_i \sim N(\delta_i, v_i)$
- It is problematic for any effect size with sparse binary outcomes to follow a normal distribution
- To avoid this part of the model, meta-analysts have used different distributions for count data for the first level

## Random Effects with Binary Data

- Many competing models have been proposed
  - Binomial – normal
  - Poisson – normal
  - Beta-binomial (non-normal conjugate distribution for the random effect)
- Still a very active area of research

## Random Effects with Binary Data

- If the binary data are **not** sparse, then random effects inverse variance methods are probably fine
- Use normalized versions of the effect sizes
  - Log odds ratio
  - Log relative risk
  - Risk difference

## Random Effects Model

- The random effects model includes an additional source of variability than the fixed effects model does, so **confidence intervals are wider**
- Random effects analysis is often considered to be more **conservative** than fixed effects
  - That is, random effects meta-analysis is considered less likely to find a statistically significant combined estimate

## Random Effects Model

- But, random effects analyses are **not always** more conservative
  - Small studies get relatively more weight in a random effects analysis than in a fixed effects analysis
  - This makes random effects analysis more affected by **small study** effects such as bias in publication
  - These biases almost always make the summary estimate go farther away from the null

## Which Model to Use?

- My approach is to do both analyses
  - Random effects as the primary analysis
  - Fixed effects as the secondary analysis
- If the two analyses are consistent, then all is fine
- If not, then need to explore why they disagree
  - This is usually due to how the small studies are treated
  - Flag for bias