STA 610L: Module 4.8

Meta-analysis

Dr. Olanrewaju Michael Akande

META-ANALYSIS

A meta-analysis is the "statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (Glass, 1976).

Meta-analysis is a standard tool for producing summaries of research findings in medicine and other fields.

Meta-analysis can be useful when studies yield potentially conflicting results, when sample sizes in individual studies are modest, when events are rare, and in general to summarize a literature.

Hierarchical models are often used as part of meta-analysis.

Example: TB studies

For our first example, we examine the results of 13 studies evaluating the efficacy of a vaccine (BCG) for preventing tuberculosis.

You can click here to see where the vaccine is given.

The vaccine is generally not recommended for use in the US due to low TB prevalence.

The data we will use in the metafor package.

This dataset has been used in several publications to illustrate meta-analytic methods.

See the documentation of the package for more details.

Example: TB studies

The goal of the meta-analysis was to examine the overall effectiveness of the BCG vaccine for preventing tuberculosis and to examine moderators that may potentially influence the size of the effect.

The data actually comes in the form of a contingency table, so we will first compute our effectiveness measure from that.

Here, we focus on log risk ratio of tuberculosis infection in the **treated versus control** groups in 13 studies.

We can also use other measures, for example, log odds ratio, if preferred.

Example: TB studies

```
#library(metafor)
data(dat.bcg)
dat.bcg
```

```
##    trial               author year tpos  tneg cpos  cneg ablat       alloc
## 1      1              Aronson 1948    4   119   11   128    44      random
## 2      2     Ferguson & Simes 1949    6   300   29   274    55      random
## 3      3      Rosenthal et al 1960    3   228   11   209    42      random
## 4      4    Hart & Sutherland 1977   62 13536  248 12619    52      random
## 5      5 Frimodt-Moller et al 1973   33  5036   47  5761    13  alternate
## 6      6      Stein & Aronson 1953  180  1361  372  1079    44  alternate
## 7      7     Vandiviere et al 1973    8  2537   10   619    19      random
## 8      8           TPT Madras 1980  505 87886  499 87892    13      random
## 9      9     Coetzee & Berjak 1968   29  7470   45  7232    27      random
## 10    10      Rosenthal et al 1961   17  1699   65  1600    42 systematic
## 11    11       Comstock et al 1974  186 50448  141 27197    18 systematic
## 12    12   Comstock & Webster 1969    5  2493    3  2338    33 systematic
## 13    13       Comstock et al 1976   27 16886   29 17825    33 systematic
```

Example: TB studies

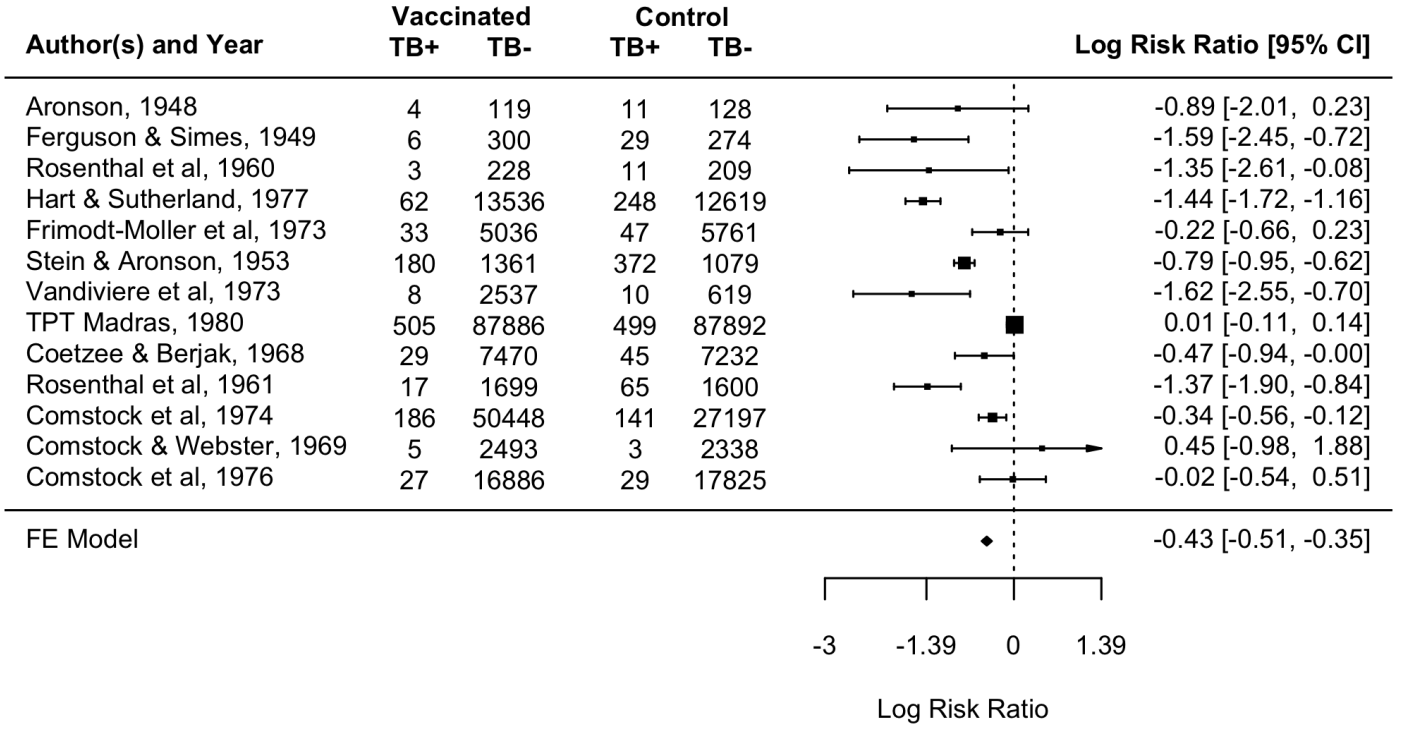
```
dat <- escalc(measure="RR", ai = tpos, bi = tneg, ci = cpos, di = cneg,
              data = dat.bcg, append = TRUE)
dat
```

```
##    trial               author year tpos  tneg cpos  cneg ablat       alloc
## 1      1              Aronson 1948    4   119   11   128    44      random
## 2      2     Ferguson & Simes 1949    6   300   29   274    55      random
## 3      3      Rosenthal et al 1960    3   228   11   209    42      random
## 4      4    Hart & Sutherland 1977   62 13536  248 12619    52      random
## 5      5 Frimodt-Moller et al 1973   33  5036   47  5761    13  alternate
## 6      6      Stein & Aronson 1953  180  1361  372  1079    44  alternate
## 7      7     Vandiviere et al 1973    8  2537   10   619    19      random
## 8      8           TPT Madras 1980  505 87886  499 87892    13      random
## 9      9     Coetzee & Berjak 1968   29  7470   45  7232    27      random
## 10    10      Rosenthal et al 1961   17  1699   65  1600    42 systematic
## 11    11       Comstock et al 1974  186 50448  141 27197    18 systematic
## 12    12   Comstock & Webster 1969    5  2493    3  2338    33 systematic
## 13    13       Comstock et al 1976   27 16886   29 17825    33 systematic
##          yi     vi
## 1  -0.8893 0.3256
## 2  -1.5854 0.1946
## 3  -1.3481 0.4154
## 4  -1.4416 0.0200
## 5  -0.2175 0.0512
## 6  -0.7861 0.0069
## 7  -1.6209 0.2230
## 8   0.0120 0.0040
## 9  -0.4694 0.0564
## 10 -1.3713 0.0730
## 11 -0.3394 0.0124
## 12  0.4459 0.5325
## 13 -0.0173 0.0714
```

Forest plot of observed values

```
res <- rma(yi, vi, data=dat, method="FE") #start with fixed effects
forest(res,
       slab = paste(dat$author, dat$year, sep = ", "),
       xlim = c(-16, 6), at = log(c(0.05, 0.25, 1, 4)),
       ilab = cbind(dat$tpos, dat$tneg, dat$cpos, dat$cneg),
       ilab.xpos = c(-9.5, -8, -6, -4.5), cex = 0.75)
op <- par(cex = 0.75, font = 2)
text(c(-9.5, -8, -6, -4.5), 15, c("TB+", "TB-", "TB+", "TB-"))
text(c(-8.75, -5.25), 16, c("Vaccinated", "Control"))
text(-16, 15, "Author(s) and Year", pos = 4)
text(6, 15, "Log Risk Ratio [95% CI]", pos = 2)
par(op)
```

Forest plot of observed values



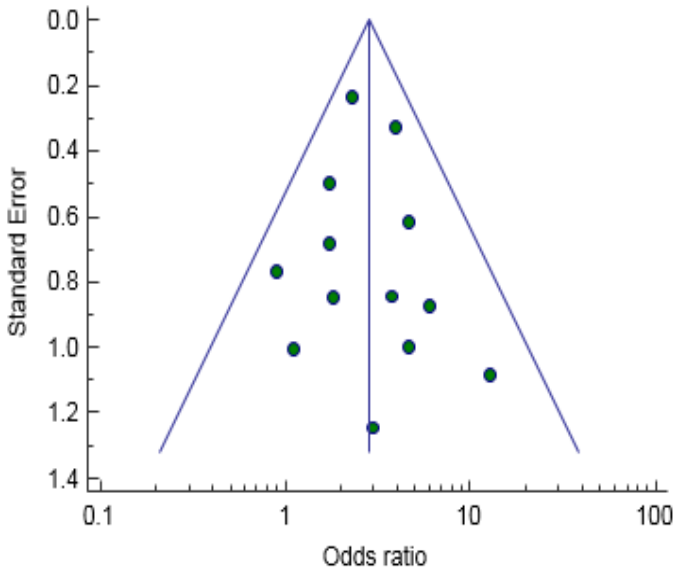
Most are below zero on the log scale and five of the confidence intervals include zero.

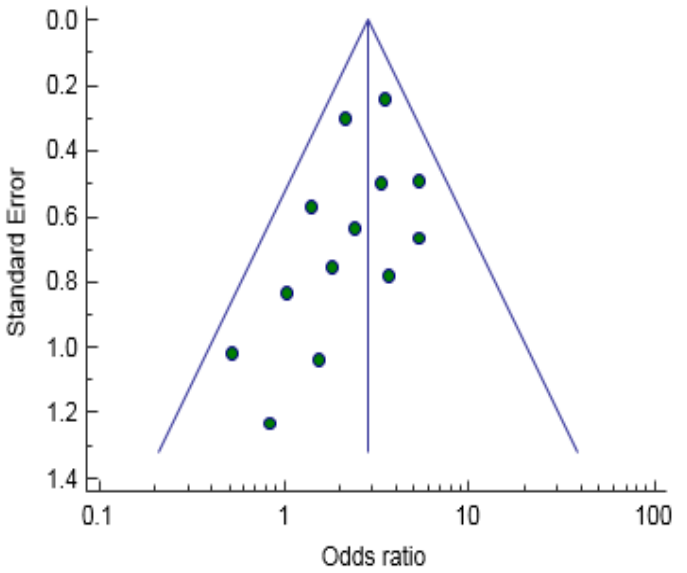
Funnel plot

Funnel plots are scatter plots of each study's effect estimates against the precision of the estimates.

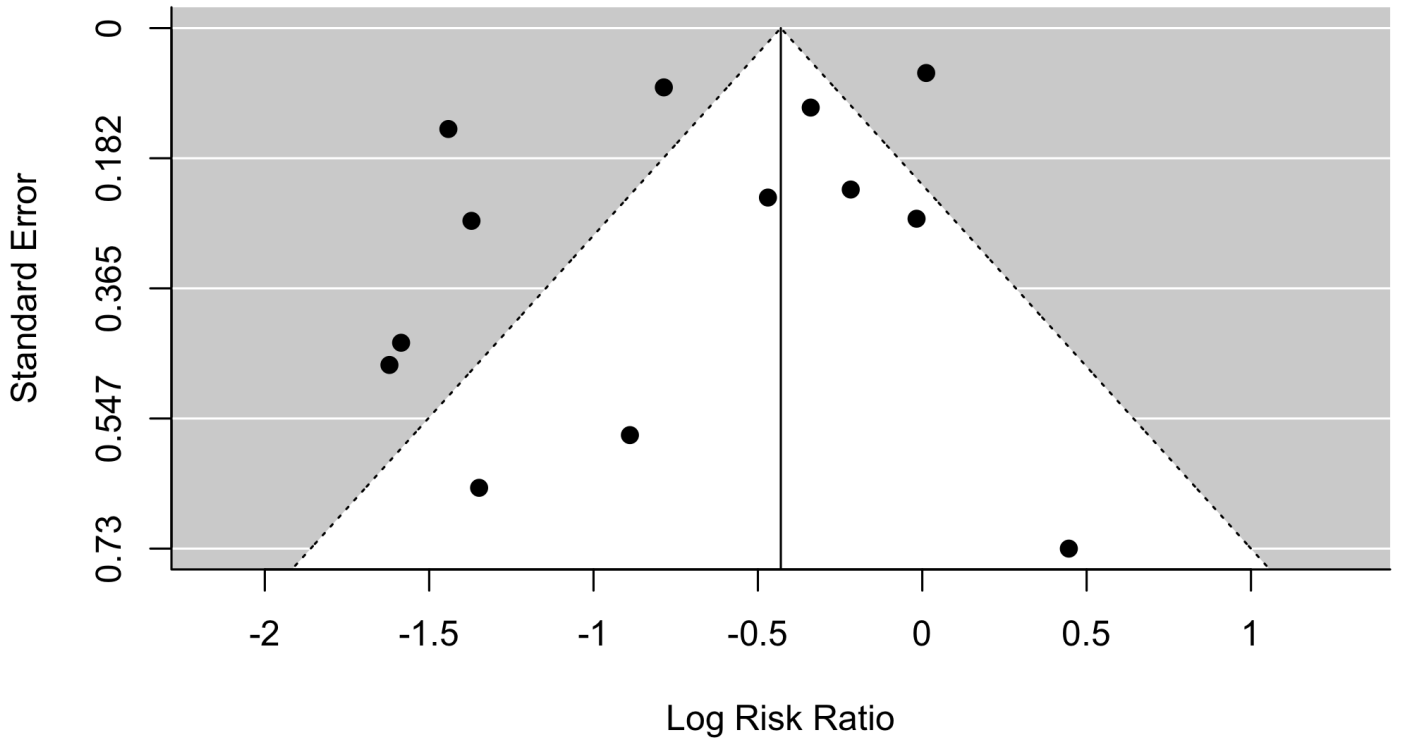
Asymmetry can indicate publication bias.

- Small, statistically insignificant studies are usually excluded from data





Funnel plot



Maybe some bias but we also see larger than expected standard errors for 6 studies.

Random effects meta-analysis

A random effects meta analysis typically assumes the model:

$$
\begin{aligned}
y_i &= \theta_i + e_i \\
\theta_i &= \mu + b_i \\
b_i &\sim N(0, \tau^2),
\end{aligned}
$$

where

- y_i is the effect estimate (possibly transformed) from study i,
- $e_i \sim N(0, v_i)$ is the sampling error from study i (the sampling variance v_i estimated from each study is assumed known),
- μ is the average "true" effect, and
- τ^2 is the heterogeneity among the study true effects.

Random effects meta-analysis

In this framework, we may think of individual studies as:

- replicates;
- results from a variety of completely different studies of the same topic;
- exchangeable yet not completely identical or unrelated.

Note the following:

- μ is typically the primary quantity of interest as a summary measure across studies;
- the error variance v_i varies across studies and is often treated as known as the square of the standard error estimate from study i.

Example: spanking data

Kurz considers data on corporal punishment of children.

UNICEF (2014) reports that 80% of children worldwide are spanked or physically punished by their parents.

Spanking is one of the most studied (and controversial) aspects of parenting, and hundreds of studies are available on the topic.

The data spank.xlsx contain 111 summary measures of a variety of child behavioral, emotional, cognitive, and physical outcomes from studies.

Example: spanking data

```
#library(readxl)
spank <- readxl::read_excel("data/spank.xlsx")
dim(spank)
```

```
## [1] 111   8
```

```
head(spank)
```

```
## # A tibble: 6 x 8
##   study                    year outcome          between within     d    ll    ul
##   <chr>                   <dbl> <chr>              <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 Bean and Roberts (198…  1981 Immediate defia…       1      0 -0.74 -1.76  0.28
## 2 Day and Roberts (1983)  1983 Immediate defia…       1      0  0.36 -1.04  1.77
## 3 Minton, Kagan, and Le…  1971 Immediate defia…       0      1  0.34 -0.09  0.76
## 4 Roberts (1988)          1988 Immediate defia…       1      0 -0.08 -1.01  0.84
## 5 Roberts and Powers (1…  1990 Immediate defia…       1      0  0.1  -0.82  1.03
## 6 Burton, Maccoby, and …  1961 Low moral inter…       0      1  0.63  0.16  1.1
```

```
length(unique(spank$outcome))
```

```
## [1] 17
```

```
length(unique(spank$study))
```

```
## [1] 76
```

Example: spanking data

```
unique(spank$outcome)
```

```
##  [1] "Immediate defiance"
##  [2] "Low moral internalization"
##  [3] "Child aggression"
##  [4] "Child antisocial behavior"
##  [5] "Child externalizing behavior problems"
##  [6] "Child internalizing behavior problems"
##  [7] "Child mental health problems"
##  [8] "Child alcohol or substance abuse"
##  [9] "Negative parent-child relationship"
## [10] "Impaired cognitive ability"
## [11] "Low self-esteem"
## [12] "Low self-regulation"
## [13] "Victim of physical abuse"
## [14] "Adult antisocial behavior"
## [15] "Adult mental health problems"
## [16] "Adult alcohol or substance abuse"
## [17] "Adult support for physical punishment"
```

Spanking data

The effect size of interest in the meta-analysis is the standardized difference in mean outcomes given by

$$d = \frac{\mu_{spanked} - \mu_{notspanked}}{\sigma_{pooled}},$$

where

$$\sigma_{pooled} = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}}.$$

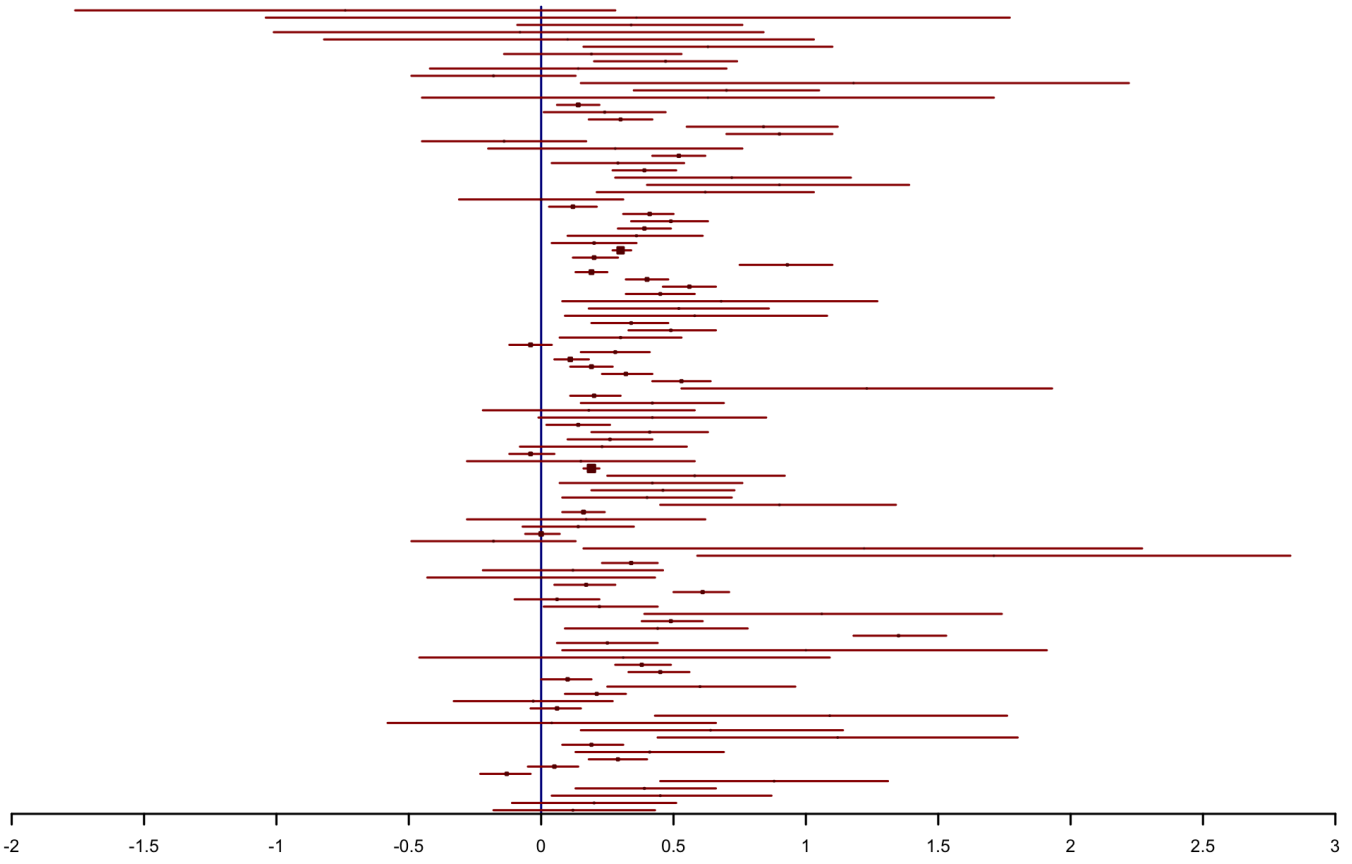
This effect size is just a mean difference converted to standard deviation units.

Most effect sizes will be fairly small -- for example, seeing an effect size of 1 would correspond to a 1 SD difference in the outcome between the spanking groups.

Let's peek at the full data in a forest plot.

Spanking data

```
forestplot(rep(NA,length(spank$study)),spank$d,spank$ll,spank$ul,
           col = fpColors(lines="#990000", box="#660000", zero = "darkblue"))
```



Spanking data

Note that the data on the previous slides do not provide us with standard errors for the effect sizes d; however, we can calculate them based on the CI's as

$$SE = \frac{\text{upper limit} - \text{lower limit}}{2 \times 1.96}.$$

```
#library(tidyverse)
spank <-
  spank %>%
  mutate(se = (ul - ll) / (2*1.96))
glimpse(spank)
```

```
## Rows: 111
## Columns: 9
## $ study   <chr> "Bean and Roberts (1981)", "Day and Roberts (1983)", "Minton, …
## $ year    <dbl> 1981, 1983, 1971, 1988, 1990, 1961, 1962, 1990, 2002, 2005, 19…
## $ outcome <chr> "Immediate defiance", "Immediate defiance", "Immediate defianc…
## $ between <dbl> 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0,…
## $ within  <dbl> 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1,…
## $ d       <dbl> -0.74, 0.36, 0.34, -0.08, 0.10, 0.63, 0.19, 0.47, 0.14, -0.18,…
## $ ll      <dbl> -1.76, -1.04, -0.09, -1.01, -0.82, 0.16, -0.14, 0.20, -0.42, -…
## $ ul      <dbl> 0.28, 1.77, 0.76, 0.84, 1.03, 1.10, 0.53, 0.74, 0.70, 0.13, 2.…
## $ se      <dbl> 0.52040816, 0.71683673, 0.21683673, 0.47193878, 0.47193878, 0.…
```

Model

$$y_i = \theta_i + e_i \quad \theta_i = \mu + b_i \quad b_i \sim N(0, \tau^2),$$

where

- y_i is the effect estimate (possibly transformed) from study i, and
- $e_i \sim N(0, v_i)$ is the sampling error from study i (the sampling variance v_i estimated from each study is assumed known).

We will go Bayesian in this example. Let's put a

- $\text{Half-Cauchy}(0, 1)$ prior on τ and
- $N(0, 1)$ prior on μ as it would be really rare to have a summary d that was very big on the effect size scale -- probably not the case for spanking but maybe if we were measuring more severe physical abuse.

Model

```
#library(brms)
m.spank <-
  brm(data = spank, family = gaussian,
      d | se(se) ~ 1 + (1 | study),
      prior = c(prior(normal(0, 1), class = Intercept),
                prior(cauchy(0, 1), class = sd)),
      iter = 4000, warmup = 1000, cores = 4, chains = 4,
      seed = 123, control = list(adapt_delta = 0.95))
```

Results

```
print(m.spank)
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: d | se(se) ~ 1 + (1 | study)
##    Data: spank (Number of observations: 111)
## Samples: 4 chains, each with iter = 4000; warmup = 1000; thin = 1;
##          total post-warmup samples = 12000
##
## Group-Level Effects:
## ~study (Number of levels: 76)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.26      0.03     0.21     0.33 1.00     1839     4066
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     0.38      0.04     0.31     0.45 1.00     1164     2476
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.00      0.00     0.00     0.00 1.00    12000    12000
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Results

Our summary measure for d is 0.38, with 95% CrI=(0.31,0.45). Kids who were spanked had on average scores 0.38 SD higher than kids who were not spanked.

These outcomes were coded by authors in the same direction, so that larger values of d imply more negative outcomes among kids who were spanked.

Note: presumably many of these studies are not randomized, and this association does not imply causation.

Multiple outcomes

One interesting aspect of the data is while we have 111 outcome effect sizes, these come from only 76 separate studies -- some studies measured multiple outcomes.

```
spank %>%
  distinct(study) %>%
  count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    76
```

We may wish to shrink outcomes of similar types together -- so let's fit a cross-classified random effects model by adding a random effect for outcome type.

Updated model

```
m.spank.outcome <-
  brm(data = spank, family = gaussian,
      d | se(se) ~ 1 + (1 | study) + (1 | outcome),
      prior = c(prior(normal(0, 1), class = Intercept),
                prior(cauchy(0, 1), class = sd)),
      iter = 4000, warmup = 1000, cores = 4, chains = 4,
      seed = 123, control = list(adapt_delta = 0.95))
```

Updated results

```
print(m.spank.outcome)
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: d | se(se) ~ 1 + (1 | study) + (1 | outcome)
##    Data: spank (Number of observations: 111)
## Samples: 4 chains, each with iter = 4000; warmup = 1000; thin = 1;
##          total post-warmup samples = 12000
##
## Group-Level Effects:
## ~outcome (Number of levels: 17)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.08      0.03     0.04     0.14 1.00     3920     6248
##
## ~study (Number of levels: 76)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.25      0.03     0.20     0.32 1.00     2976     5059
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     0.36      0.04     0.28     0.44 1.00     2950     4853
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.00      0.00     0.00     0.00 1.00    12000    12000
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

The estimates of d are quite similar to our previous ones. Looking at the variance components, the study-to-study heterogeneity is larger than heterogeneity across outcomes. We can explore further in a figure.

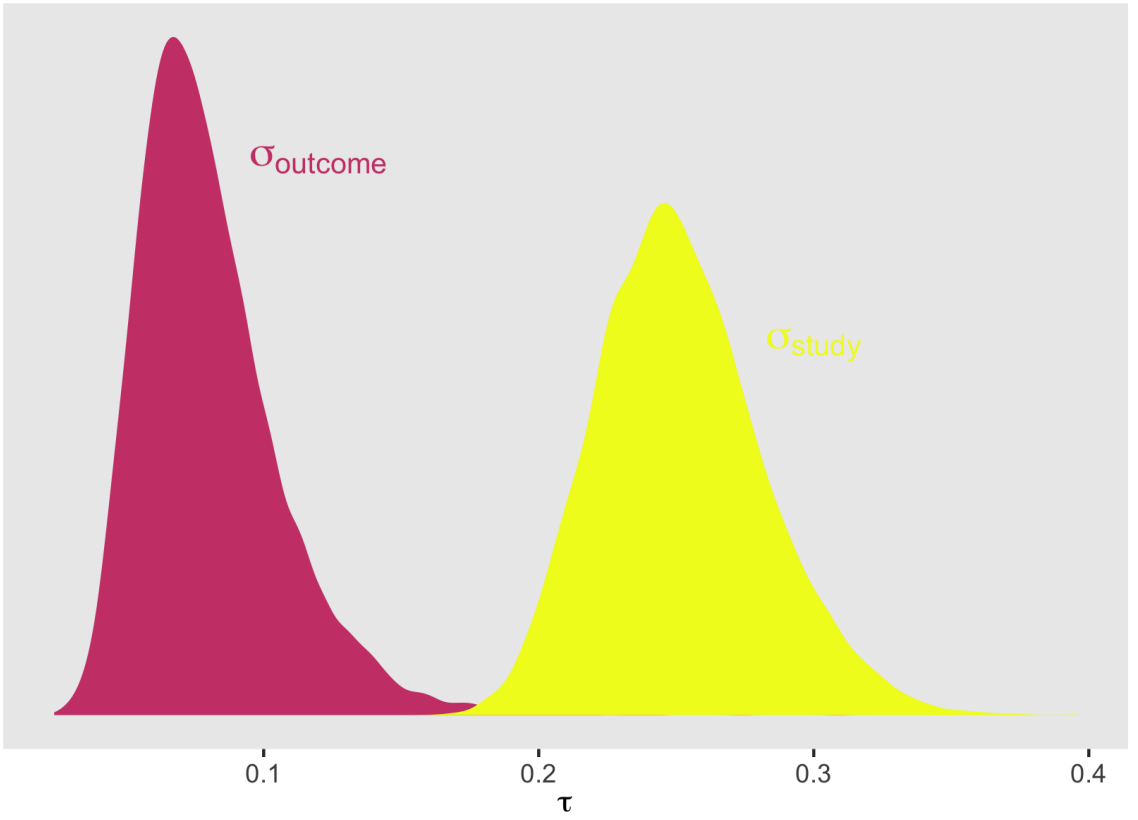
Updated results

```
# we'll want this to label the plot
label <-
  tibble(tau   = c(.12, .3),
         y     = c(15, 10),
         label = c("sigma['outcome']", "sigma['study']"))

# wrangle
posterior_samples(m.spank.outcome) %>%
  select(starts_with("sd")) %>%
  gather(key, tau) %>%
  mutate(key = str_remove(key, "sd_") %>% str_remove(., "__Intercept")) %>%

  # plot
  ggplot(aes(x = tau)) +
  geom_density(aes(fill = key),
               color = "transparent") +
  geom_text(data = label,
            aes(y = y, label = label, color = label),
            parse = T, size = 5) +
  scale_fill_viridis_d(NULL, option = "B", begin = .5) +
  scale_color_viridis_d(NULL, option = "B", begin = .5) +
  scale_y_continuous(NULL, breaks = NULL) +
  xlab(expression(tau)) +
  theme(panel.grid = element_blank())
```

Updated results



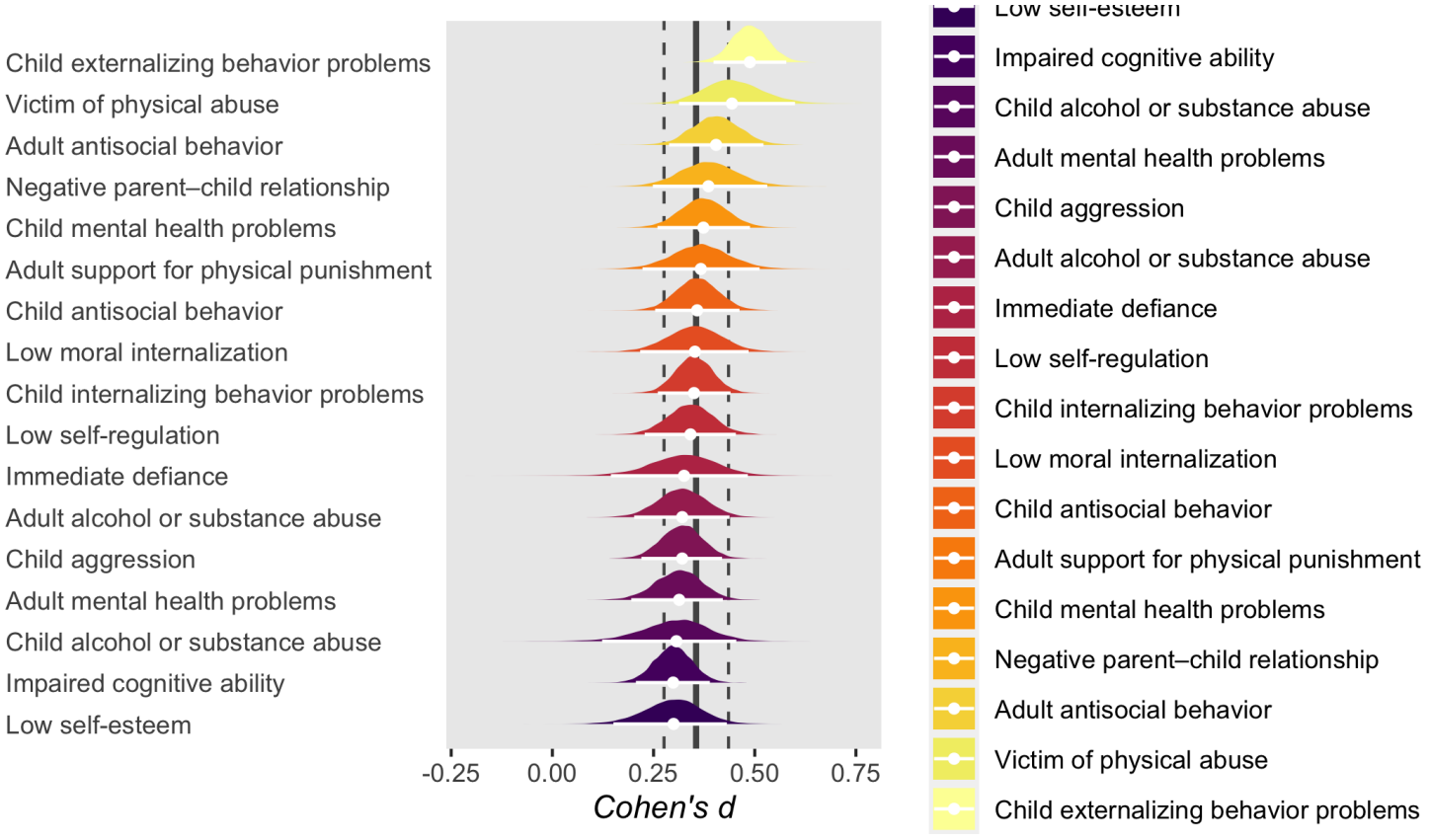
Updated results

We can also check whether spanking has similar effects on all the different outcomes -- let's examine those more closely.

```
#library(tidybayes)
m.spank.outcome %>%
  spread_draws(b_Intercept, r_outcome[outcome,]) %>%
  # add the grand mean to the group-specific deviations
  mutate(mu = b_Intercept + r_outcome) %>%
  ungroup() %>%
  mutate(outcome = str_replace_all(outcome, "[.]", " ")) %>%

  # plot
  ggplot(aes(x = mu, y = reorder(outcome, mu),
             fill = reorder(outcome, mu))) +
  geom_vline(xintercept = fixef(m.spank.outcome)[1, 1],
             color = "grey33", size = 1) +
  geom_vline(xintercept = fixef(m.spank.outcome)[1, 3:4],
             color = "grey33", linetype = 2) +
  geom_halfeyeh(.width = .95, size = 2/3, color = "white") +
  scale_fill_viridis_d(option = "B", begin = .2) +
  labs(x = expression(italic("Cohen's d")),
       y = NULL) +
  theme(panel.grid   = element_blank(),
        axis.ticks.y = element_blank(),
        axis.text.y  = element_text(hjust = 0))
```

Updated results



We see evidence that spanking may be particularly linked with child externalizing behavior problems (again, this is chicken & egg -- we cannot infer causation).

Example: bladder cancer

There are many other interesting variations of this standard random effects model.

For example, we may want to assign weights to the studies, especially when we do not have that many studies to work with, and we think the studies vary in quality.

In our next example, we have results from seven studies about the effect of chlorinated water on the odds ratio of getting bladder cancer.

Five studies investigated a sample cancer deaths, while two studies looked at cancer diagnoses.

There is likely natural (or maybe systematic) variability across these studies.

Example: bladder cancer

Again, the goal is to combine the results of these studies to estimate the "true" overall effect, incorporating information about the quality of study and uncertainty of estimates of effect size.

Author	Year	AdjOR	LCL	UCL	Method	Quality
Cantor	1987	1.19	1.07	1.32	Logistic	78
Zierler	1988	1.60	1.20	2.10	M-H	71
Wilkins	1986	2.20	0.71	6.82	Logistic	61
Gottlieb	1982	1.18	0.95	1.45	Adj	49
Brenniman	1980	0.98	0.77	1.25	Adj	46
Young	1981	1.15	0.70	1.89	Logistic	45
Alvanja	1978	1.69	1.07	2.67	Adj	43

Example: bladder cancer

```
author <- c("Cantor","Zierler","Wilkins","Gottlieb","Brenniman", "Young", "Alvanja")
year <- c(1987, 1988, 1986, 1982, 1980, 1981, 1978)
adjOR <- c(1.19, 1.60, 2.20, 1.18, .98, 1.15, 1.69)
LCL <- c(1.07, 1.2, .71, .95, .77, .7, 1.07)
UCL <- c(1.32, 2.10, 6.82, 1.45, 1.25, 1.89, 2.67)
method <- c("Logistic", "M-H", "Logistic", "Adj", "Adj", "Logistic", "Adj")
quality <- c(78, 71, 61, 49, 46, 45, 43)

meta <- data.frame(author, year, adjOR, LCL, UCL, method, quality)

#convert to log odds ratio so we can use a linear mixed effects model
meta$LN_adjOR <- round(log(meta$adjOR),2)

#also get the standard error on the log odds ratio scale
meta$SE_LNadjOR <- round((log(meta$UCL) - log(meta$adjOR))/1.96,2)
```

Example: bladder cancer

Note: M-H is the Mantel-Haenszel method, which produces and approximate logistic regression estimate.

The odds ratio was adjusted by some method other than logistic regression.

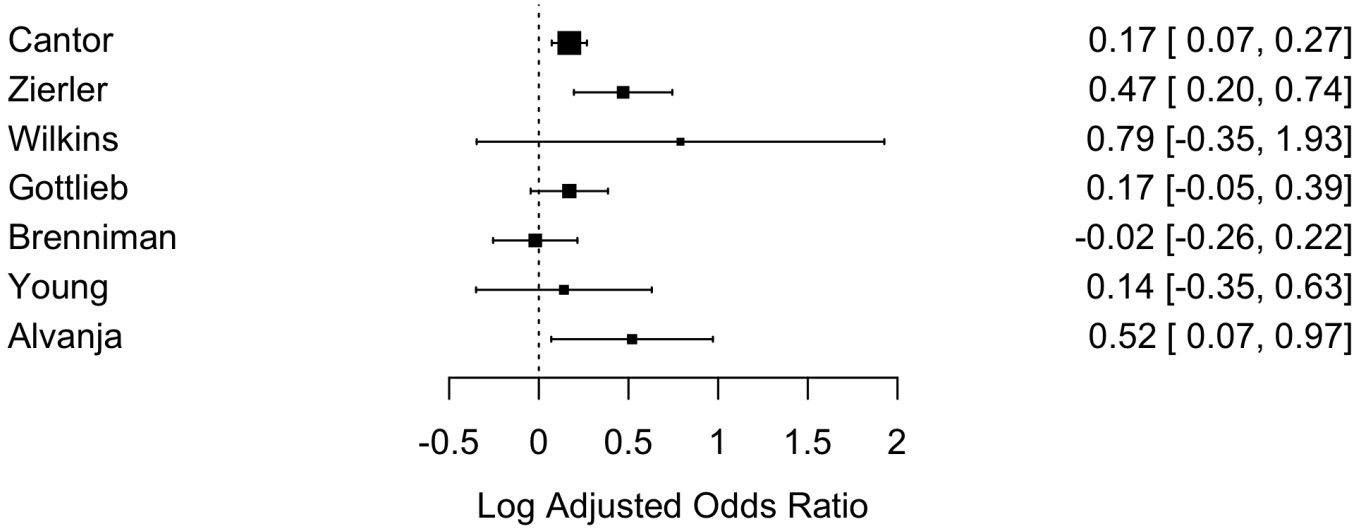
Each paper was rated for quality on the basis of selection of subjects, measurement of and adjustment for confounding variables, exposure assessment, and statistical analysis.

Interpret the score as the percentage of quality.

Easy to think about weighting each study using a function of its quality rating.

Forest plot

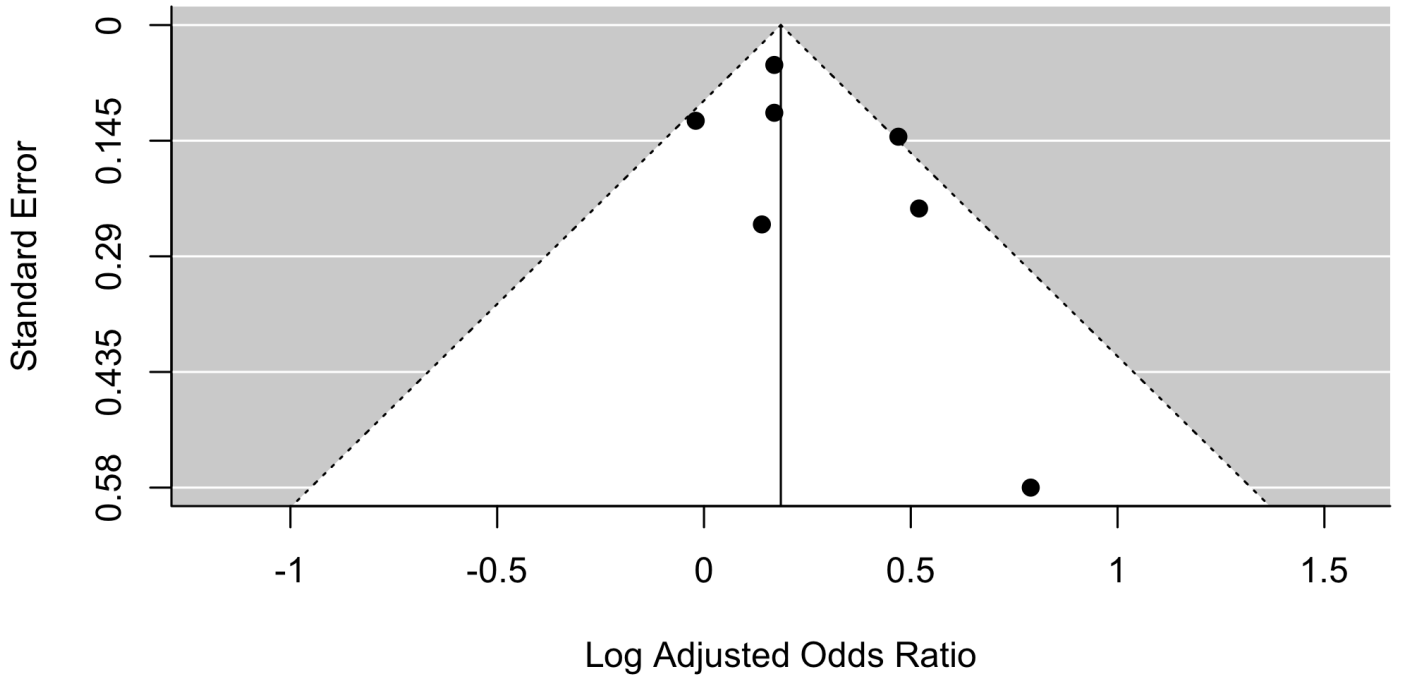
```
forest(x=meta$LN_adjOR, sei=meta$SE_LNadjOR, slab=meta$author, top=0.5,
       xlab="Log Adjusted Odds Ratio")
```



All log-odds ratio estimates are above zero, with the exception of Brenniman.

Four of the seven confidence intervals include zero.

Funnel plot



No immediate publication bias seems evident in the data. Difficult to determine asymmetry in the plot because there are only seven studies.

Random effects model with weights

Suppose

- y_i is the log odds ratio for study i, and
- w_i is the weight given to study i.

Then we can fit the following model

$$y_i = \theta_i + e_i; \quad \theta_i = \mu + b_i$$
$$b_i \sim N(0, \tau^2); \quad e_i \sim N(0, v_i),$$

and estimate the overall effect as

$$\hat{\mu} = \frac{\sum_i w_i y_i}{\sum_i w_i}; \quad \text{with} \quad Var(\hat{\mu}) = \frac{\sum_i w_i^2 Var(y_i)}{(\sum_i w_i)^2}.$$

Some options for the weights

The weights should obviously be related to the model but how should we specify them? Here are some common options:

- Option I: $w_i = \frac{1}{\tau^2 + v_i}$
  - Each study is weighted by the sample variance with more weight on studies with lower variance
- Option II: $w_i = Q_i$
  - Each study is weighted by quality with more weight on studies with higher quality.
- Option III: $w_i = \frac{\hat{Q}_i}{\tau_i^2 + v_i}$
  - $\hat{Q}_i$ is a modified quality measure, with more weight on studies with high quality and low variance

The variances are estimated from the random effects model. Note: the second option does not require any model.

Quality effects model for meta-analysis

Option III incorporates quality by adjusting the weight as well as redistributing weights based on quality. (Doi, Thalib, 2009).

Note:

- Q_i is quality of study i
- N is total number of studies.

Then, we have

$$w_i = \frac{1}{\tau^2 + v_i^2} \qquad \tau_i = \frac{w_i - (w_i \cdot Q_i)}{N - 1}$$

$$\hat{\tau}_i = \sum_i \tau_i - \tau_i \quad \textit{is a quality adjustor}$$

$$\hat{Q}_i = Q_i + \frac{\hat{\tau}_i}{w_i} \quad \textit{is the modified quality.}$$

Final comments

Easy to implement all three options, especially using the metafor package.

This is a very short introduction to meta-analysis in R but is as much as we are going to cover.

The metafor package allows for many kinds of models for meta-analysis.

When fitting Bayesian version, also use the brms package as always.

For a much more detailed material on meta-analysis (both classicial and Bayesian), see this very wonderful hands-on guide!

Also, take a look at Section 15.5 of A. Solomon Kurz's statistical rethinking ebook.

What's next?

Move on to the readings for the next module!