Pre-registration of 'Between-Study Variation in Meta-Analyses of Mechanical Thrombectomy: A Systematic Review and Case Study of Statistical Reporting Practices in Clinical Medicine'

Bradley Kolb

Rationale

Between-study variation is essential for interpreting the average treatment effect reported in a meta-analysis, as it quantifies how sensitive the treatment effect is to the clinical context. When variation is high – due to differences in patient populations, protocols, facility expertise, or other factors – treatment effects in specific settings may differ substantially from meta-analytic averages, regardless of the average effect size or its statistical significance.

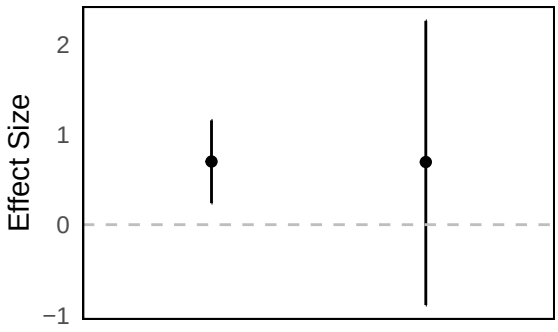


Figure 1: The illusion of certainty in meta-analysis: a significant average effect (left) does not imply a positive trial-specific effect (right).

In an effort to better understand how well the current medical literature communicates between-study variation in randomized trials, we propose to systematically review reporting

practices for between-study variation in meta-analyses of mechanical thrombectomy for acute ischemic stroke. This literature offers an ideal case study for multiple reasons:

- implementation characteristics
  - well-defined intervention
  - measurable uniform outcomes
  - identifiable effect modifiers
- evidence-base
  - multiple high-quality RCTs
  - enables robust estimation of meta-analytic parameters
- clinical impact
  - published meta-analyses directly influence clinician behavior and clinical guidelines
- publication characteristics
  - multiple meta-analyses
  - diverse journal types
  - diverse author perspectives

By analyzing how these meta-analyses handle between-study variation, we aim to assess current reporting practices and develop recommendations that better support evidence-based decision-making.

Background

Formally, a random-effects meta-analysis is a multilevel model with the following structure.

$$y_j \sim \mathrm{N}(\theta_j, \sigma_j)$$
$$\theta_j \sim \mathrm{N}(\mu, \tau)$$

From a generative perspective, this model says that the true underlying treatment effect θ_j for a trial j is a function of the average treatment effect μ and the between-trial variation τ, and that the effect we actually observe in j is a random deviation from θ_j governed by σ_j, the standard error of y in j.

From a statistical perspective, what this model says is that we can combine the observed information

$$(y_1, \sigma_1)...(y_n, \sigma_n)$$

in n clinical trials to estimate both the true underlying treatment effects for each trial

$$\theta_1...\theta_n$$

as well as the average treatment effect across the population of all trials μ and the between-trial variation across the population τ.

Although μ has a more intuitive interpretation ("the average treatment effect"), τ is arguably the more important parameter in the random-effects model, because variation in treatment effects across trials reveals how context-dependent the intervention is. While μ estimates the average effect (the most likely outcome in any single context), τ quantifies the range of plausible effects (what effects could plausibly be observed in any single context). Together, these parameters tell us not just how effective a treatment is on average, but how reliably that effectiveness translates across settings. Importantly, a precise estimate for μ ("significantly" greater than zero, say) *does not* imply that the effect in any one trial will be greater than zero. To know what the range of plausible outcomes is in any given trial, you need to know τ.

Estimating τ faces two key challenges:

1. Statistical complexity
    - Reliable estimation of τ requires modern advanced algorithms (Bayesian MCMC or frequentist optimization, for instance)
    - Common meta-analytic software (e.g., RevMan) uses approximation formulas that can be inaccurate
2. Communication barriers
    - τ lacks the (seemingly) intuitive interpretation of μ ("average effect")
    - many analyses bypass reporting τ entirely, in favor of statistics such as I^2 or Q

These challenges likely affect reporting quality, but their impact hasn't been systematically evaluated in a robust clinical literature. This study seeks to close this gap by examining how meta-analyses of mechanical thrombectomy trials report between-study variation.

By analyzing how between-study variation is currently reported, we aim to:

- Assess whether clinicians receive adequate information about treatment effect variability
- Identify specific practices that impede proper interpretation of between-study variation in meta-analyses

- Develop recommendations for improved reporting standards that better support evidence-based decision-making, which critically requires an understanding of between-study variation

Inclusion criteria

1. Population and intervention
    - Articles that perform a meta-analysis on the efficacy or safety of mechanical thrombectomy for acute ischemic stroke.
2. Publication sources and time frame
    - Journals: JAMA, The Lancet, JAMA Neurology, Neurology, Stroke, Journal of Neurointerventional Surgery, American Journal of Neuroradiology, and Neurosurgery
    - Publication dates: January 1, 2014 to January 1, 2024
3. Study design
    - Must self-identify as meta-analysis in the title
    - Must apply a random effects model
4. Outcome measures
    - Primary or secondary outcomes relevant to mechanical thrombectomy efficacy or safety
5. Exclusion criteria
    - Meta-analyses not using a random-effects model
    - Narrative reviews, systematic reviews without quantitative meta-analysis, or single-study analyses

Search strategy

1. Databases

    - Pubmed
    - Hand search to augment and check results

2. Search term

    - ((((((((((“Lancet (London, England)”[Journal]) OR (“The Lancet. Neurology”[Journal])) OR (“JAMA”[Journal])) OR (“jama neurology”[Journal])) OR (“Neurology”[Journal])) OR (“Stroke”[Journal])) OR (“AJNR. American journal of neuroradiology”[Journal])) OR (“Journal of neurointerventional surgery”[Journal])) OR (“Neurosurgery”[Journal])) AND (meta-analysis[Title])) AND (thrombectomy[Title])) AND ((“2014”[Date - Publication] : “2024”[Date - Publication]))seq

3. Time filter

    - 2014/01/01 to 2024/12/31.

4. Screening process

    - Titles and abstracts will be screened by one reviewer to identify potential meta-analyses of mechanical thrombectomy.
    - Full texts of eligible citations will be retrieved, and full-text eligibility screening will confirm that a random-effects model was used.

5. Documentation

    - The number of articles identified, screened, included, and excluded will be reported in a PRISMA flow diagram.

Data extraction

A standardized data extraction form will be developed. We will pilot the form on 5 randomly selected articles and then iteratively refine it.

General study information

- pmid, article title, first author, journal, year

Meta-analysis characteristics

- whether the statistical software package used to perform the analysis was reported (if open source software such as R or python was used, must specify packages used, if closed-source software such as RevMan, SPSS, etc, then specifying that fact is sufficient)
- if so, what package was used
- Whether code was shared
- Whether enough information was reported to repeat the analysis
- Whether any model checking was performed (for instance, how sensitive were conclusions to model specification choices? to choice of random effects model over fixed effects model? etc)
- Whether a statistician was one of the authors or was consulted

Problematic practices in reporting between-trial heterogeneity

1. Complete omission

    - no reporting of heterogeneity statistics in body of article
    - fails to address fundamental aim of random-effects meta-analysis

2. Inadequate reporting of τ

    - the fundamental measure of between-trial heterogeneity
    - in a pilot review of 5 randomly selected studies, τ was not reported in the body of any of the studies

3. Misuse of I^2

    - Incorrect definition: I^2 represents the proportion of total variance due to between-trial heterogeneity. It is often reported incorrectly as a measure of heterogeneity itself
    - Even if I^2 is defined correctly, its interpretation is scale-dependent, insofar as it depends on the underlying value for τ
    - In our pilot review, I^2 was not defined or incorrectly defined in all 5 studies, was never interpreted with respect to the underlying τ estimate, and was often reported as exact zero without qualification
    - but reporting $I^2 = 0$ without qualification is problematic, because it reflects statistical limitations rather than true empircal homogeneity
    - Other problematic misuses of I^2 observed in our pilot studies included applying arbitrary thresholds to I^2 values and inappropriately treating I^2 as a test statistic

4. Misuse of Q and χ^2

    - in general, applying null hypothesis testing to the question of between-trial variation is discouraged
    - however, even when choosing to employ this potentially problematic practice, interpreting non-significant p values as evidence of homogeneity is a statistical fallacy

- nonetheless, in our pilot study, this practice was observed multiple times

5. Poor integration with clinical context

    - reporting measures without discussing implications
    - not making connection to interpretation of μ (external validity)
    - not using prediction intervals

Planned analyses

1. Descriptive statistics
2. Interpretation

    - narrative interpretation of results
    - implication of results for utility of published meta-analyses
    - recommendations for improved reporting standards
    - discussion of specific problematic practices

3. Re-analysis

    - For each paper, identify the main analysis, and if it reports enough data, replicate the meta-analysis using Stan and two methods
        - full posterior distributions using MCMC and regularizing priors (bayesian)
        - penalized maximum likelihood point-estimates using optimization (frequentist)
    - compare these results to reported parameter estimates
    - hypothesis: the re-analyzed estimates for τ and I^2 will be different than many of the reported point-estimates in the literature (because these estimates use the RevMan "DerSimonian and Laird random-effects model" to obtain the estimates, which defaults to $\tau = 0$ when Cochrane's Q is less than or equal to its degrees of freedom$k - 1$.
    - implication: many meta-analyses would be misrepresenting the degree of between-trial heterogeneity in the literature

Technical details

Simulating clinical trial data sets in R

The following function does simulates a single data-set of randomized trials assuming a true underlying value for μ and a true underlying value for τ.

The following function uses `sim_one_dataset` to simulate multiple data sets of multiple different sizes.

For example, we can use `sim_many_datasets` to produce 100 sets of 3 simulated trials, 100 sets of 5 simulated trials, and 100 sets of 20 simulated trials, all generated from the same underlying values for μ and τ.

With these data sets in hand, we than apply various estimation methods to see which are best able to recover the underlying true parameter values, and to assess how these techniques perform when the number of simulated trials varies.

Implementing standard random effects meta-analysis in R

The Dersimonian-Laird estimator for the random effects meta-analysis model is implemented as follows.

We can generate one set of simulated trials and use the DL method to estimate μ, τ, and I^2.

```
True values:

tau: 0.7

mu: 0.7

average within-study standard error: 0.2129264


Results from DerSimonian-Laird Method:

Estimated tau (heterogeneity): 0.6066311

Random-effects summary effect size: 0.7038684

Standard error of summary effect size: 0.203355
```

```
I^2: 89.20957
```

Check correctness using the `metafor` package in `R`.

```
Random-Effects Model (k = 10; tau^2 estimator: DL)

tau^2 (estimated amount of total heterogeneity): 0.3680 (SE = 0.1960)
tau (square root of estimated tau^2 value):      0.6066
I^2 (total heterogeneity / total variability):   89.21%
H^2 (total variability / sampling variability):  9.27

Test for Heterogeneity:
Q(df = 9) = 83.4072, p-val < .0001

Model Results:

estimate      se    zval    pval   ci.lb   ci.ub
  0.7039  0.2034  3.4613  0.0005  0.3053  1.1024  ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When the number of studies in a meta-analysis is high, this approach performs relatively well. When the number of studies is low, it performs less well. We can show this graphically and numerically as follows.

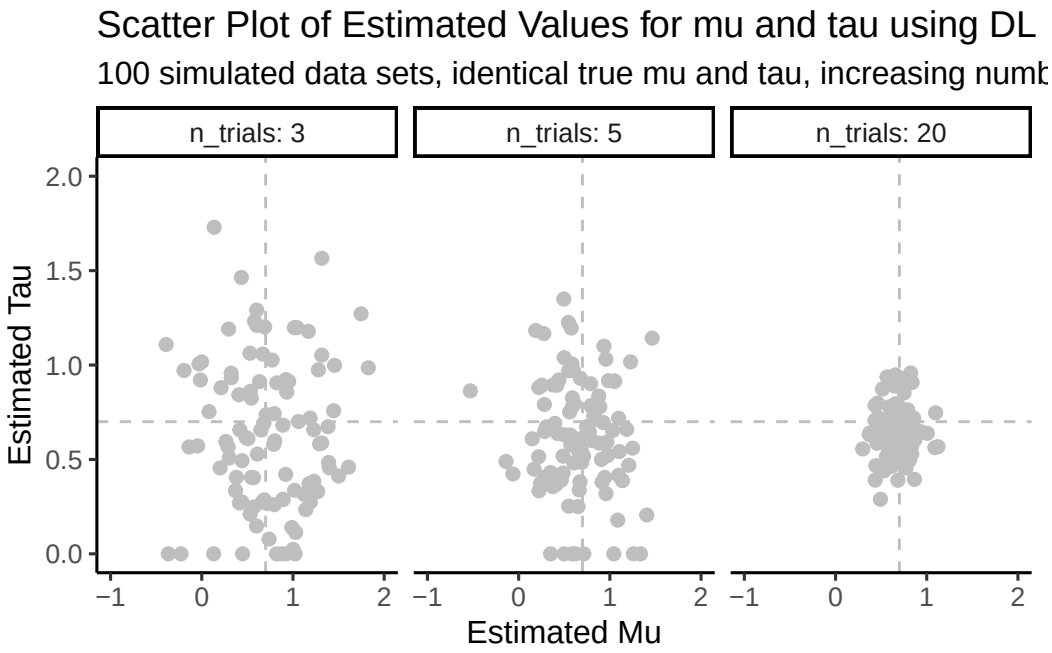
First, we will write a generic function that applies an estimator to a data-set.

Next, we will wrap this function so it can be re-used.

Next, we will simulate a data set.

Finally, we will apply the estimator function to the simulated data set.

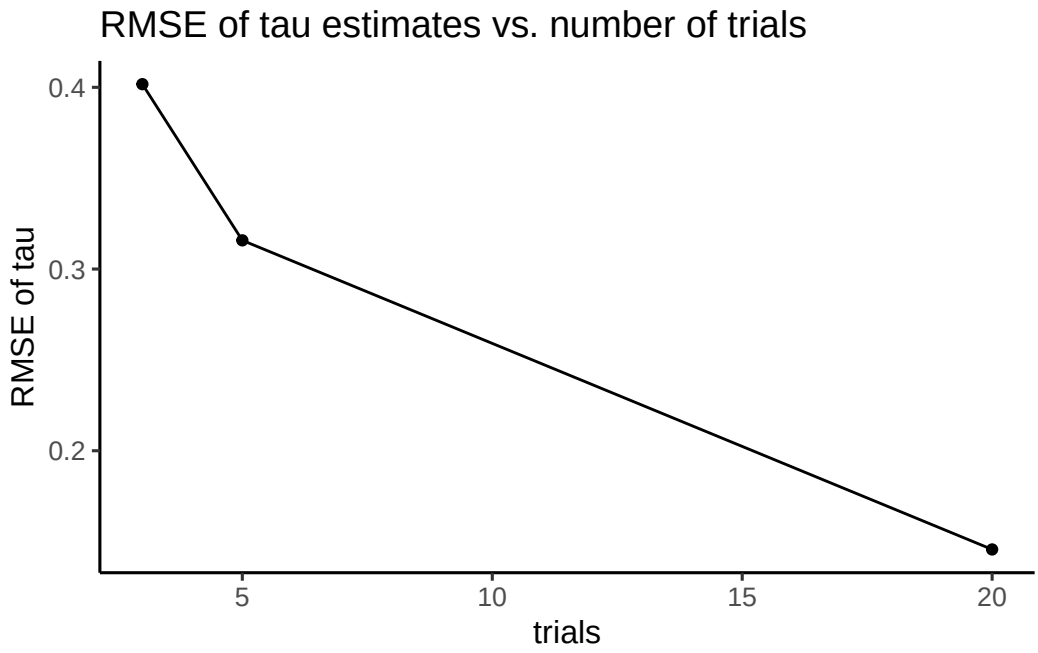
Then, we can graph the results.



As the number of studies increases from 3 to 5 to 20, the estimates for μ and τ concentrate more closesly around the true underlying values of 0.7 for each. In other words, the accuracy of the DL method improves greately with increased study number. We can show this numerically by calculating the root mean squared error (RMSE) for each of the 3 sets of simulated data.

```
# A tibble: 3 x 3
  n_trials rmse_mu rmse_tau
     <dbl>   <dbl>    <dbl>
1        3   0.469    0.402
2        5   0.346    0.316
3       20   0.155    0.146
```

We can also graph these results.



Estimating between-trial variation using maximum likelihood estimation

We can implement the basic random-effects meta-analysis model in `Stan`.

We can obtain the maximum-likelihood estimates for the parameter values using `Stan`'s built-in optimizer.

```
Initial log joint probability = -266.543
    Iter      log prob        ||dx||      ||grad||       alpha      alpha0  # evals  Notes
      13       -0.4004   1.12191e-05   7.88423e-05       0.966       0.966       15
Optimization terminated normally:
  Convergence detected: relative gradient magnitude is below tolerance
Finished in  0.1 seconds.
```

```
# A tibble: 3 x 2
  variable estimate
  <chr>       <dbl>
1 mu          0.703
2 tau         0.580
3 i2          0.880
```

Estimating between-trial variation using Bayesian inference

We can implement a Bayesian version of random-effects meta-analysis in `Stan`.

We can obtain the estimated posterior distribution for the model using MCMC sampling.

```
Running MCMC with 4 parallel chains...

Chain 1 finished in 0.3 seconds.
Chain 2 finished in 0.3 seconds.
Chain 3 finished in 0.2 seconds.
Chain 4 finished in 0.3 seconds.

All 4 chains finished successfully.
Mean chain execution time: 0.3 seconds.
Total execution time: 0.5 seconds.


# A tibble: 3 x 10
  variable  mean median     sd    mad    q5   q95  rhat ess_bulk ess_tail
  <chr>    <dbl>  <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl>    <dbl>    <dbl>
1 mu       0.670  0.672 0.238  0.220  0.271 1.05   1.01     599.     849.
2 tau      0.705  0.669 0.197  0.176  0.444 1.08   1.00     681.     969.
3 i2       0.900  0.907 0.0466 0.0434 0.812 0.962  1.00     681.     969.
```

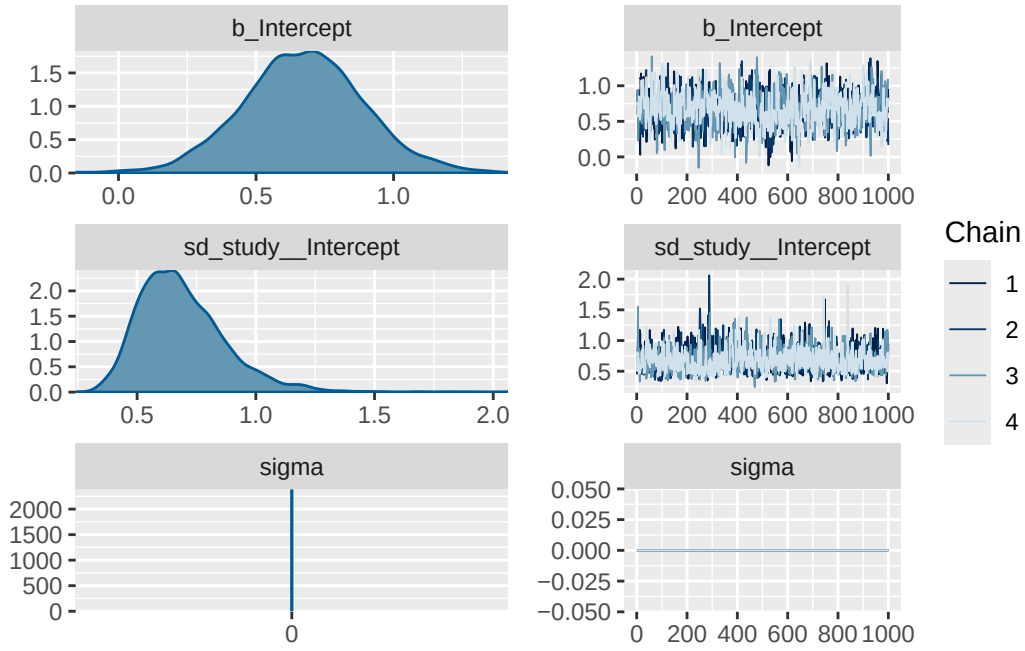
Compare to results of Bayesian model using `brms`.

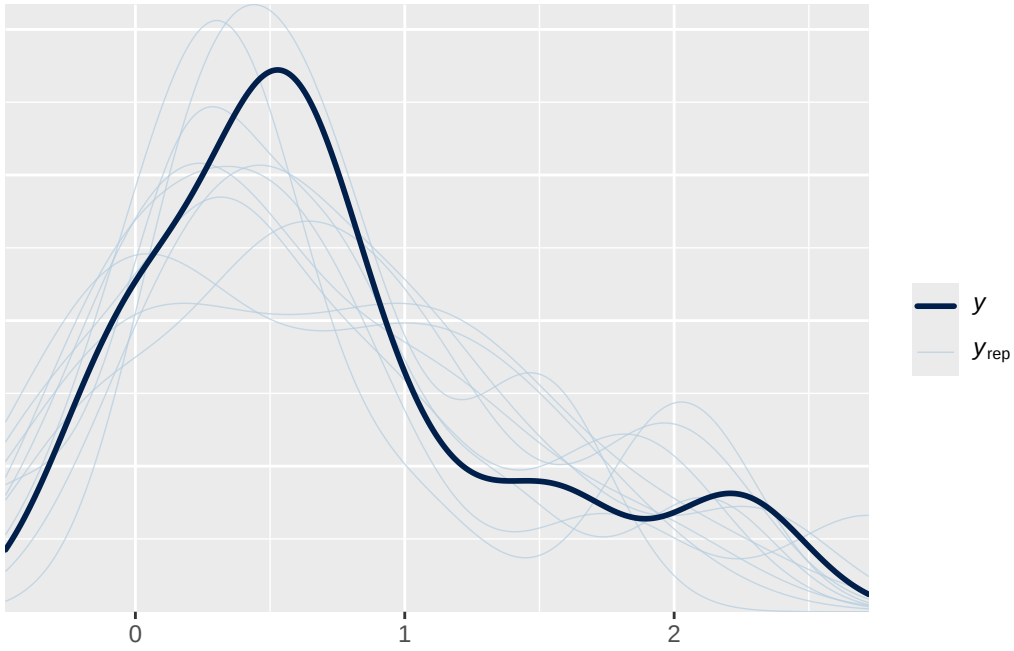
```
Running MCMC with 4 parallel chains...

Chain 1 Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 1 Iteration:  100 / 2000 [  5%]  (Warmup)
Chain 1 Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 1 Iteration:  300 / 2000 [ 15%]  (Warmup)
Chain 1 Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 1 Iteration:  500 / 2000 [ 25%]  (Warmup)
Chain 1 Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 1 Iteration:  700 / 2000 [ 35%]  (Warmup)
Chain 1 Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 1 Iteration:  900 / 2000 [ 45%]  (Warmup)
Chain 1 Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 1 Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 1 Iteration: 1100 / 2000 [ 55%]  (Sampling)
Chain 1 Iteration: 1200 / 2000 [ 60%]  (Sampling)
```

```
Chain 1 Iteration: 1300 / 2000 [ 65%]  (Sampling)
Chain 1 Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 1 Iteration: 1500 / 2000 [ 75%]  (Sampling)
Chain 1 Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 1 Iteration: 1700 / 2000 [ 85%]  (Sampling)
Chain 2 Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 2 Iteration:  100 / 2000 [  5%]  (Warmup)
Chain 2 Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 2 Iteration:  300 / 2000 [ 15%]  (Warmup)
Chain 2 Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 2 Iteration:  500 / 2000 [ 25%]  (Warmup)
Chain 2 Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 2 Iteration:  700 / 2000 [ 35%]  (Warmup)
Chain 2 Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 2 Iteration:  900 / 2000 [ 45%]  (Warmup)
Chain 2 Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 2 Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 2 Iteration: 1100 / 2000 [ 55%]  (Sampling)
Chain 2 Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 2 Iteration: 1300 / 2000 [ 65%]  (Sampling)
Chain 2 Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 2 Iteration: 1500 / 2000 [ 75%]  (Sampling)
Chain 2 Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 2 Iteration: 1700 / 2000 [ 85%]  (Sampling)
Chain 2 Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 3 Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 3 Iteration:  100 / 2000 [  5%]  (Warmup)
Chain 3 Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 3 Iteration:  300 / 2000 [ 15%]  (Warmup)
Chain 3 Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 3 Iteration:  500 / 2000 [ 25%]  (Warmup)
Chain 3 Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 3 Iteration:  700 / 2000 [ 35%]  (Warmup)
Chain 3 Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 3 Iteration:  900 / 2000 [ 45%]  (Warmup)
Chain 3 Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 3 Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 3 Iteration: 1100 / 2000 [ 55%]  (Sampling)
Chain 3 Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 3 Iteration: 1300 / 2000 [ 65%]  (Sampling)
Chain 3 Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 3 Iteration: 1500 / 2000 [ 75%]  (Sampling)
Chain 3 Iteration: 1600 / 2000 [ 80%]  (Sampling)
```

```
Chain 3 Iteration: 1700 / 2000 [ 85%]  (Sampling)
Chain 3 Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 3 Iteration: 1900 / 2000 [ 95%]  (Sampling)
Chain 3 Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 4 Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 4 Iteration:  100 / 2000 [  5%]  (Warmup)
Chain 4 Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 4 Iteration:  300 / 2000 [ 15%]  (Warmup)
Chain 4 Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 4 Iteration:  500 / 2000 [ 25%]  (Warmup)
Chain 4 Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 4 Iteration:  700 / 2000 [ 35%]  (Warmup)
Chain 4 Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 4 Iteration:  900 / 2000 [ 45%]  (Warmup)
Chain 4 Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 4 Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 4 Iteration: 1100 / 2000 [ 55%]  (Sampling)
Chain 4 Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 4 Iteration: 1300 / 2000 [ 65%]  (Sampling)
Chain 4 Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 4 Iteration: 1500 / 2000 [ 75%]  (Sampling)
Chain 4 Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 4 Iteration: 1700 / 2000 [ 85%]  (Sampling)
Chain 4 Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 4 Iteration: 1900 / 2000 [ 95%]  (Sampling)
Chain 1 Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 1 Iteration: 1900 / 2000 [ 95%]  (Sampling)
Chain 1 Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 1 finished in 0.1 seconds.
Chain 2 Iteration: 1900 / 2000 [ 95%]  (Sampling)
Chain 2 Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 2 finished in 0.1 seconds.
Chain 3 finished in 0.1 seconds.
Chain 4 Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 4 finished in 0.1 seconds.

All 4 chains finished successfully.
Mean chain execution time: 0.1 seconds.
Total execution time: 0.5 seconds.


 Family: gaussian
  Links: mu = identity; sigma = identity
```

```
Formula: observed_effects | se(standard_errors) ~ 1 + (1 | study)
   Data: one_dataset (Number of observations: 10)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Group-Level Effects:
~study (Number of levels: 10)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.69      0.18     0.41     1.13 1.00     1075     1541

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     0.67      0.22     0.24     1.12 1.00      915     1516

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   NA       NA       NA

Draws were sampled using sample(hmc). 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).
```




0
1
2
y
y_rep