To appear in Statistical Methods in Medical Research

Meta-analysis with a single study

Erik van Zwet[*], Witold Więcek[†], Andrew Gelman[‡]

December 19, 2024

Abstract

Effect sizes typically vary among studies of the same intervention. In a random effects meta-analysis, this source of variation is properly taken into account. However, when we have only one study, the heterogeneity remains hidden and unaccounted for. Treating the study-level effect as if it is the population-level effect leads to underestimation of the uncertainty. We propose an empirical Bayesian approach to address this problem. We start by estimating the distribution of the population-level effects and heterogeneity among 1,635 meta-analyses from the Cochrane Database of Systematic Reviews (CDSR). Using both synthetic data and cross-validation, we assess the consequences of using these estimated distributions as prior information for the analysis of single trials. We find that our Bayesian "meta-analyses of single studies" perform much better than naively assuming non-varying effects. The prior on the heterogeneity results in better quantification of the uncertainty. The prior on the treatment effect substantially reduces the mean squared error both for estimating the study-level and population-level effects. For the latter, this reduction is equivalent to doubling the sample size.

1 Introduction

It is a common practice to interpret the observed treatment effect of a clinical trial as an estimate of the underlying "true" treatment effect. This is of course

[*] Department of Biomedical Data Sciences, Leiden University Medical Center, The Netherlands

[†] Development Innovation Lab, University of Chicago, USA

[‡] Department of Statistics and Department of Political Science, Columbia University, USA

quite reasonable. We claim that it is also common to tacitly, perhaps even unconsciously, assume that this true effect is an immutable property of the treatment. Thus, one might speak of “the” effect of a particular treatment or intervention, and expect the same effect in another study about the same treatment. However, there are many good reasons to expect variation among the underlying effects from differences between study populations, application of the treatment and measurement protocols among other factors. There is also much empirical evidence of effect heterogeneity from random effects meta-analyses [8, 11, 7]. Failure to take this into account when interpreting the results of trial will lead to underestimation of the uncertainty about the treatment effect.

Recall that meta-analysis is a quantitative approach for combining inferences from multiple studies. Besides providing an estimate of the average effect of the treatment, together with some measure of uncertainty such as a standard error or confidence interval, a meta-analysis also provides an estimate of the variation of underlying effects of the individual studies. In fact, understanding and quantifying heterogeneity is an important aspect of meta-analysis, as it can influence the interpretation of the overall results.

The Cochrane Database of Systematic Reviews (CDSR) is a globally respected collection of evidence-based healthcare information, comprising rigorous and comprehensive systematic reviews on diverse medical topics, with many of those reviews including meta-analyses [4]. The database is continuously updated, offering the latest evidence to inform clinical practice, policy decisions, and research priorities. It adheres to strict quality standards, undergoes peer review, and includes open-access summaries for wider accessibility.

We will assume the standard random effects meta-analysis model. That is, we assume the following two-part hierarchical (or multilevel) model for the j-th study in a collection of studies of the same treatment

$$\beta_j = \mu + u_j \tag{1}$$

$$b_j = \beta_j + \varepsilon_j, \tag{2}$$

The error terms $u_j \sim \text{normal}(0, \tau)$ and $\varepsilon_j \sim \text{normal}(0, s_j)$ are assumed to be independent. The first part (1) models the variation among studies of the same treatment. Here, μ is the population-level average treatment effect in the (hypothetical) superpopulation of similar studies and β_j is the study-level effect in the j-th individual study. The variance τ^2 is referred to as the heterogeneity. The second part (1) models the uncertainty of the estimates from each individual study. We assume that the estimate b_j from the j-th study is unbiased and normally distributed with standard error s_j.

When we have only one study, we cannot separate the two error terms. Without additional assumptions, there is nothing more to do than estimate β_1 by b_1 with confidence interval $b_1 \pm 1.96\, s_1$. If we make the extra assumption that there is no heterogeneity, i.e. $\tau = 0$, then we can also estimate the population average effect μ by b_1 with confidence interval $b_1 \pm 1.96\, s_1$. As we discussed, we believe that this extra assumption is commonly—if implicitly—made when interpreting the result of a single trial. However, τ is typically not zero, so this is not correct.

Instead of assuming that τ is zero, we propose a Bayesian approach with the prior information based on a hierarchical model fit to the CDSR. First, we estimate the distributions of both μ and τ across the CDSR and use those as empirical priors. Estimating one or both of these distributions from collections of meta-analysis is not new, see for instance [6, 8, 11]. We then proceed to use the `R` package `baggr` [13] to use these distributions as prior information for Bayesian inference. We call this meta-analysis with a single trial. The `baggr` package is based on Stan, a state-of-the-art platform for statistical modeling and high-performance statistical computation [10].

We compare the performance across the trials of the CDSR of our Bayesian approach to the naive approach of assuming that τ is zero. To do this, we construct a "synthetic copy" of the CDSR and also use a kind of leave-one-out cross validation. We find superior performance of the Bayesian approach on average across the CDSR.

2 Meta-analysis with a single trial

2.1 Estimating the distributions of μ and τ

The primary results of the trials in the CDSR have been processed and made available by [9]. We use these to estimate the typical variation between the effects of the same treatment across multiple studies.

We start by selecting trials with either a binary or numerical primary efficacy outcome; these comprise 97% of the trials in the CDSR. To make the effect sizes of the binary and numerical outcomes comparable, we quantify the treatment effect on the probit scale for all binary outcomes, and as the standardized mean difference (SMD) for all continuous outcomes. Next, we select all meta-analyses with at least 5 individual studies. Meta-analyses with fewer studies have little information about heterogeneity and discarding them reduces computation time. This leaves 18,368 unique trials from 1,635 meta-analyses. We consider the following hierarchical model. For the j-th individual study in the i-th meta-analysis, we assume

	t distribution of μ			normal of $\log\tau$	
	center	scale	df	mean	std dev
unrestricted	-0.11	0.37	5.30	-1.82	0.89
zero mean	0.00	0.37	5.18	-1.82	0.90

Table 1: *Estimated parameters of the t distribution of μ and the normal distribution of* $\log \tau$. *In the bottom row, we restrict the mean of μ to be zero.*

$$\beta_{ij} = \mu_i + u_{ij} \tag{3}$$

$$b_{ij} = \beta_{ij} + \varepsilon_{ij} \tag{4}$$

where $u_{ij} \sim \text{normal}(0, \tau_i)$ and $\varepsilon_{ij} \sim \text{normal}(0, s_{ij})$. All the u_{ij} and ε_{ij} are assumed to be independent. From the CDSR we obtain the pairs (b_{ij}, s_{ij}). We define the z-values for each pair as $z_{ij} = b_{ij}/s_{ij}$.

We want to estimate the distribution of the μ_i and τ_i across the CDSR. We assume that the μ_i follow a generalized t distribution and that the τ_i are lognormally distributed. Moreover, we assume that the μ_i and τ_i are independent.

To estimate the 5 parameters of our model (the mean, scale and degrees of freedom of the generalized t-distribution and the mean and standard deviation of the normal distribution) we use a Bayesian approach with uniform priors on the two means, the degrees of freedom, and the logarithm of the two scale parameters. The posterior distributions of the parameters are approximately normal, so that the posterior means are approximately equal to the maximum likelihood estimates. We show these estimates in the top row of Table 1.

We can provide some context for these estimated distributions by recalling the tentative classification of effect sizes by Cohen according to which SMD values of 0.2 to 0.5 are considered small, 0.5 to 0.8 are considered medium, and greater than 0.8 are considered large [3, 1]. The estimated t-distribution of μ implies that the median of the absolute value of μ is 0.28 (IQR: 0.13 to 0.50). In other words, 75% of the population-level average effects may be considered to be small. The median of τ is 0.16 (IQR: 0.09 to 0.30). So, we find that the heterogeneity is roughly on the order of half the effect size.

We also fit the model restricting the center of the distribution of effect sizes μ_i to be zero, and show the resulting estimates in the bottom row of Table 1. We will use this distribution as our prior, because it ensures that we treat positive and negative effect estimates equally. This seems fair because, to some extent, the sign of the effect estimate is arbitrary. For example, one could either consider the proportion of patients alive or dead after one year, but this choice should not have any material effect on our inferences about the effectiveness of the treatment.

2.2 An example

We demonstrate our approach with a small example. Suppose that we have a single trial with a numerical outcome, with estimated SMD of $b = 0.7$. Suppose that the standard error also 0.7, so that the z-value is 1 and the p-value is 0.32. The 95% confidence interval is $(-0.7, 2.1)$.

We can use the `R` package `baggr` to incorporate the prior information that is represented in Table 1. We first use a flat prior for the population-level effect μ and an informative prior for the heterogeneity.

```
bg <- baggr(data.frame(tau = 0.7, se = 0.7),
            prior_hypersd = lognormal(-1.82, 0.90))
```

We find that the posterior distribution of the study-level effect (the true effect in the trial) β remains numerically the same. That is, the posterior mean is $\hat{\beta} = 0.7$ with 95% posterior interval from -0.7 to 2.1. Now we also obtain an estimate of the population-level effect (the average effect in similar trials) μ. The posterior mean is $\hat{\mu} = 0.7$ with 95% posterior interval $(-0.9, 2.2)$. The prior information about the heterogeneity is reflected in a wider interval.

Next, we also incorporate information about the population-level effect μ. We use the following call to `baggr`.

```
bg <- baggr(data.frame(tau = 0.7, se = 0.7),
            prior_hypermean = student_t(5.18, 0, 0.37),
            prior_hypersd = lognormal(-1.82, 0.90))
```

The zero-mean prior for μ induces considerable shrinkage. We find that the posterior mean of β is now $\hat{\beta} = 0.24$ with 95% posterior interval from -0.5 to 1.1. The posterior mean of μ becomes $\hat{\mu} = 0.17$ with 95% posterior interval $(-0.5, 1.0)$. Moreover, using the code from the Appendix, we find that the posterior probability that β is positive is 0.72, while the posterior probability that μ is positive is 0.67.

2.3 The probability of the correct sign

Next, we use `baggr` to perform all 18,368 meta-analyses with one trial and compute the posterior probabilities that the observed b_{ij} has the same sign as β_{ij} and μ_i. We plot the observed absolute z-values $|z_{ij}| = |b_{ij}/s_{ij}|$ versus these probabilities in Figure 1. We fitted smooth regression curves which can be interpreted as conditional probabilities given the observed absolute z-value. We see that a single trial never provides certainty about the sign of μ.

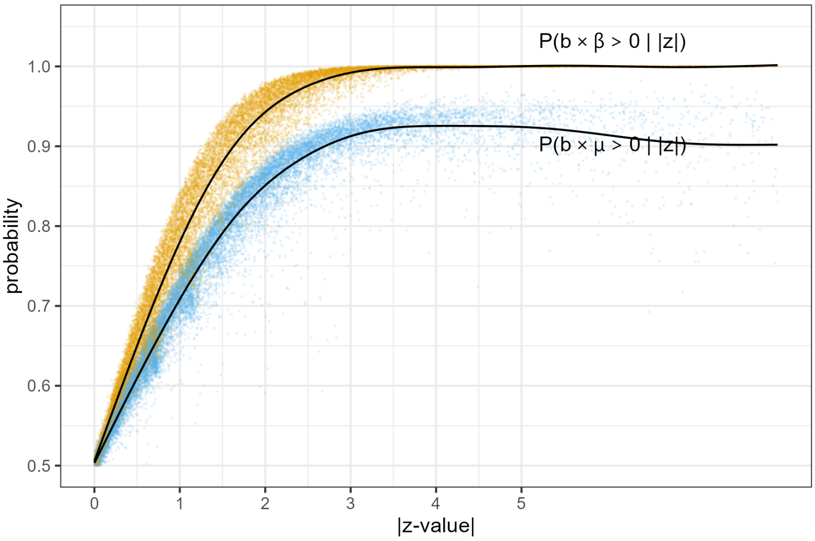

Figure 1: *Conditional probabilities of b having the same sign as β or μ given the observed absolute z-value.*

3 Performance of the method

3.1 Building a synthetic CDSR

We construct a "synthetic" CDSR to evaluate the performance of our Bayesian approach and compare it to the naive approach of assuming that τ is zero. To generate the synthetic database we perform the following steps:

1. Sample μ_i^* and τ_i^* $(i = 1, 2, \ldots, 1635)$ from the estimated distribution in the top row of Table 1. We are not restricting the μ^* to be zero on average.

2. To induce dependence between the pairs μ_i^* and τ_i^*, sort them in the same order as the maximum likelihood estimates of μ_i and τ_i, which we obtain by performing the meta-analyses of the CDSR data. Break any ties at random.

3. Sample independent β^*_{ij} ($i = 1, 2, \ldots, 1635$ and $j = 1, 2, \ldots, n_i$) from the normal distribution with mean μ^*_i and standard deviation τ^*_i.
4. Sample b^*_{ij} from the normal distribution with mean β^*_{ij} and the *observed* standard deviation s_{ij}. Set $z^*_{ij} = b^*_{ij}/s_{ij}$.

For $i = 1, 2, \ldots, 1635$ and $j = 1, 2, \ldots, n_i$, we now have simulated sets (μ^*_i, τ^*_i) and $(\beta^*_{ij}, b^*_{ij}, s_{ij})$ that should be similar to the original CDSR. We can confirm this to some extent by comparing the distribution of the observed b_{ij} and z_{ij} to the simulated b^*_{ij} and z^*_{ij} in Figure 2.

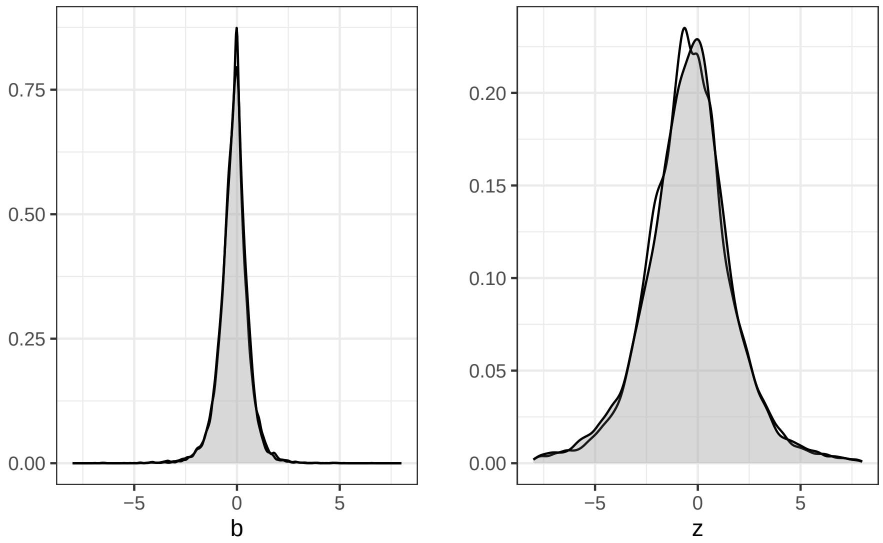

Figure 2: *The observed and simulated distributions of 18,368 estimates and z-values.*

3.2 Estimating the effect in the trial

Once again we use `baggr` to perform all meta-analyses with one trial in the simulated dataset to estimate β^*_{ij} and μ^*_{ij} by their respective posterior means $\hat{\beta}^*_{ij}$ and $\hat{\mu}^*_{ij}$. We also construct uncertainty ("credible") limits. We then compare to the naive approach where we estimate both quantities by b^*_{ij} with interval $(b^*_{ij} \pm 1.96 \times s_{ij})$.

In Table 2 we show the root mean squared error (RMSE), bias of the magnitude and coverage of these two estimators. To be precise, for the naive approach we compute

$$\text{mean squared error} = \frac{1}{18368}\sum_{i,j}(b^*_{ij} - \beta^*_{ij})^2 \tag{5}$$

$$\text{bias of the magnitude} = \frac{1}{18368}\sum_{i,j}|b^*_{ij}| - |\beta^*_{ij}| \tag{6}$$

$$\text{coverage} = \frac{1}{18368}\sum_{i,j}\mathbf{1}\{|b^*_{ij} - \beta^*_{ij}| < 1.96 s_{ij}\} \tag{7}$$

We compute the analogous quantities for the Bayesian approach. The left side Table 2 displays the three performance measures for all trials and the right side by averaging only over the statistically significant trials with $|z^*_{ij}| > 1.96$.

	unconditional			statistically significant		
method	RMSE	bias	coverage	RMSE	bias	coverage
unbiased	0.38	0.10	0.95	0.42	0.21	0.90
Bayes	0.29	-0.06	0.95	0.29	-0.01	0.95

Table 2: *Performance of the unbiased and Bayesian estimators of the effect in the trial β. The right side of the table shows the average over the statistically significant trials, i.e. $|z^*_{ij}| > 1.96$.*

Over all trials, the bias of the magnitude for the naive estimator is 0.1, which is due to Jensen's inequality; recall that the absolute value is a convex function. As expected, the coverage of the usual confidence interval equals its nominal level. The right side of the table shows that selection on significance increases the upward bias of the magnitude to 0.21. This is sometimes called the "winner's curse." This bias also causes the mean squared error to increase. Moreover, the usual confidence interval no longer reaches nominal coverage.

When we turn to our Bayesian approach, we find that the RMSE is substantially reduced compared to the unbiased estimator. In fact, the MSE is reduced from $0.38^2 = 0.14$ to $0.29^2 = 0.08$, which would be equivalent to an increase of the sample size of more than 71%! Conditionally on statistical significance, the reduction of the RMSE is even greater.

The reductions in the RMSE are due to the "shrinkage" that is induced by the zero-mean prior for μ which implies a zero-mean prior for β. This pulls the unbiased naive estimate towards zero, i.e. $|\hat{\beta}^*_{ij}| < |b^*_{ij}|$. We also see the effect of shrinkage in the reduction of the bias of the magnitude. On average across all the trials, the bias of the magnitude of the Bayesian estimator is -0.06. When we condition on statistical significance, however, this bias is almost exactly offset by the winner's curse. The coverage of the Bayesian uncertainty interval is nominal, both unconditionally and conditionally. The superior performance of a similar Bayesian estimator was seen in [12].

3.3 Estimating the population average effect

We also have two estimators for the population average effect μ_i^*, namely the naive estimator b_{ij}^* and the Bayesian estimator $\hat{\mu}_{ij}^*$. Again, we compute the root mean squared error (RMSE), bias of the magnitude and coverage both conditionally and unconditionally on statistical significance. We show the results in Table 3.

Essentially, the same observations apply to the results Table 3 as to those in Table 2. The reduction in the RMSE of the Bayesian estimator compared to the unbiased estimator is even more extreme. The MSE is reduced by more than a factor 2 from $0.53^2 = 0.28$ $0.35^2 = 0.12$ which is equivalent to more than doubling the sample size. Also, the bias of $|b_{ij}^*|$ as an estimator of $|\mu_{ij}^*|$ is large, especially for statistically significant trials. This bias is completely absent for the Bayesian estimator.

The coverage of the usual confidence interval is far below 95% both with and without conditioning on statistical significance. This is direct result of not taking the heterogeneity into account. The coverage of the Bayesian uncertainty interval is close to nominal in both cases.

	unconditional			statistically significant		
method	RMSE	bias	coverage	RMSE	bias	coverage
unbiased	0.53	0.15	0.82	0.72	0.41	0.63
Bayes	0.35	-0.11	0.93	0.38	-0.03	0.94

Table 3: *Performance of the unbiased and Bayesian estimators of the population average effect μ. The right side of the table shows the average over the statistically significant trials, i.e. $|z_{ij}^*| > 1.96$.*

3.4 Graphical comparison

Tables 2 and 3 provide a broad overview of the performance of the naive and Bayesian estimators. We will now study the performance in some more detail both from the frequentist and Bayesian points of view. The frequentist point of view means that we condition on the true effects β_{ij}^* and μ_i^*. The Bayesian point of view, on the other hand, means that we condition on the observed effect b_{ij}^*.

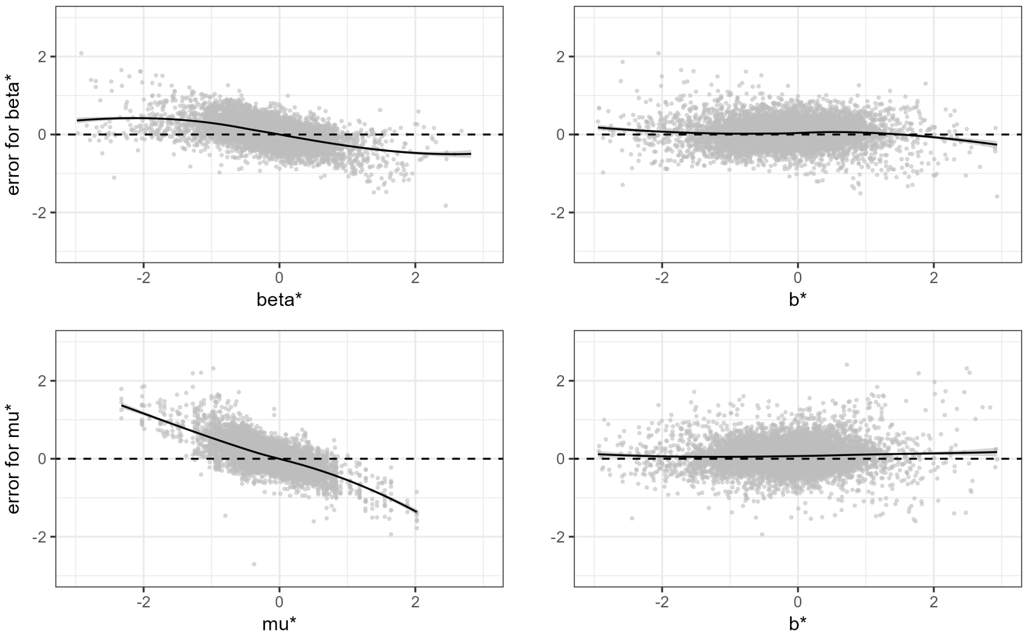

Figure 3: *The bias of the proposed Bayesian estimators. The two top panels show $\hat{\beta}^*_{ij} - \beta^*_{ij}$. The two bottom panels show $\hat{\mu}^*_{ij} - \mu^*_i$. The two panels on the left side show the frequentist bias, while the two panels on the right side show the Bayesian bias.*

We first consider the bias. Of course, the naive estimator b^*_{ij} is unbiased both for β^*_{ij} and μ^*_i so we do not need to study this further. We focus on the bias of the Bayesian estimators $\hat{\beta}^*_{ij}$ and $\hat{\mu}^*_{ij}$ which are both shrinkage estimators, in the sense that $|\hat{\mu}^*_{ij}| < |\hat{\beta}^*_{ij}| < |b^*_{ij}|$.

In the top left panel of Figure 3 we plot the estimation errors $\hat{\beta}^*_{ij} - \beta^*_{ij}$ versus the true effects in the trials β^*_{ij}. In the bottom left panel we plot the errors $\hat{\mu}^*_{ij} - \mu^*_i$ versus the true pooled effects μ^*_{ij} versus. In the two right panels, we plot the same estimation errors, but in both cases we put the observed effects b^*_{ij} on the x-axis. We added the loess regression curves to each of the 4 plots, which estimate the following conditional expectations

$$\begin{aligned}
&\mathbb{E}(\hat{\beta}^*_{ij} - \beta^*_{ij} \mid \beta^*_{ij}) && \text{(top left panel of Figure 3)}\\
&\mathbb{E}(\hat{\mu}^*_{ij} - \mu^*_i \mid \mu^*_i) && \text{(bottom left panel of Figure 3)}\\
&\mathbb{E}(\hat{\beta}^*_{ij} - \beta^*_{ij} \mid b^*_{ij}) && \text{(top right panel of Figure 3)}\\
&\mathbb{E}(\hat{\mu}^*_{ij} - \mu^*_i \mid b^*_{ij}) && \text{(bottom right panel of Figure 3)}
\end{aligned}$$

The Bayesian estimators $\hat{\beta}^*_{ij}$ and $\hat{\mu}^*_{ij}$ are both biased in the frequentist sense,

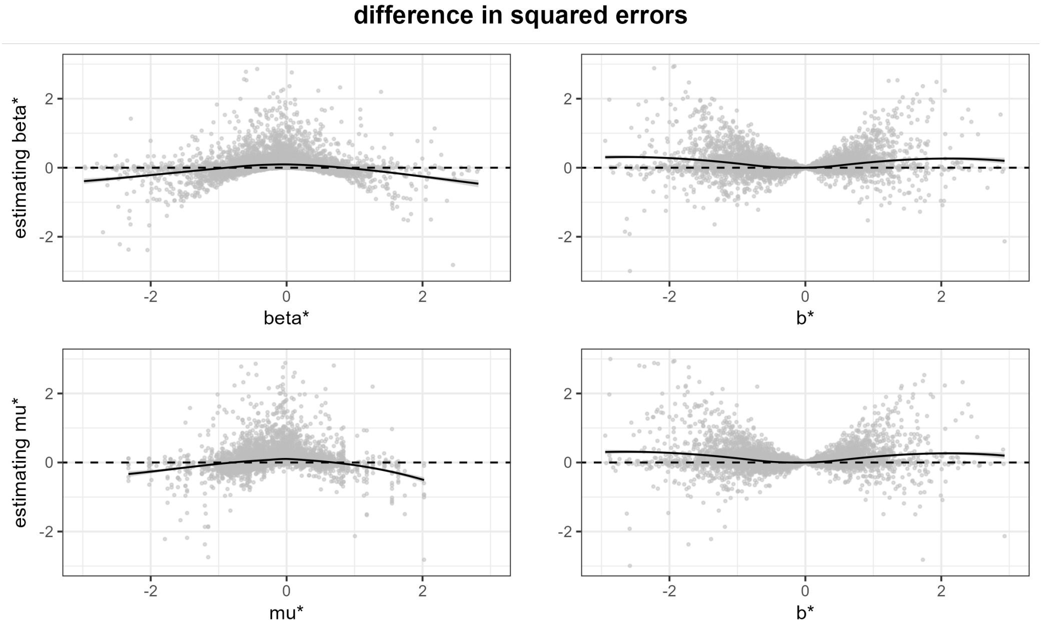

Figure 4: *Difference in squared error of the naive and Bayesian estimators. Top row: difference in squared errors for β^*_{ij} from (8). Bottom row: difference in squared errors for μ^*_i, from (9). The two panels on the left side show the frequentist perspective, while the two panels on the right side show the Bayesian perspective.*

that is, conditional on the true parameter value. This bias is apparent in the two left panels Figure 3.

The two right panels paint a different picture. They show the bias in the Bayesian sense, that is, conditional on the observed effect. In this sense, the bias is negligible! The small bias that remains is due to our choice of a zero-mean prior while the average of the b^*_{ij} is slightly negative.

Figure 4 shows the difference of the squared errors between the naive and Bayesian estimators. To be specific, in the top row we show these differences for estimating the β^*_{ij}

$$(b^*_{ij} - \beta^*_{ij})^2 - (\hat{\beta}^*_{ij} - \beta^*_{ij})^2 \tag{8}$$

and in the bottom row we show them for estimating the μ^*_i

$$(b^*_{ij} - \mu^*_i)^2 - (\hat{\mu}^*_{ij} - \mu^*_i)^2. \tag{9}$$

In both cases, positive values favor the Bayesian estimators. Again, we added the loess regression curves, which now estimate the following conditional expectations

$$
\begin{array}{ll}
\mathbb{E}((b^*_{ij}-\beta^*_{ij})^2-(\hat\beta^*_{ij}-\beta^*_{ij})^2 \mid \beta^*_{ij}) & \text{(top left panel of Figure 4)} \\
\mathbb{E}((b^*_{ij}-\mu^*_{i})^2-(\hat\mu^*_{ij}-\mu^*_{i})^2 \mid \mu^*_{i}) & \text{(bottom left panel of Figure 4)} \\
\mathbb{E}((b^*_{ij}-\beta^*_{ij})^2-(\hat\beta^*_{ij}-\beta^*_{ij})^2 \mid b^*_{ij}) & \text{(top right panel of Figure 4)} \\
\mathbb{E}((b^*_{ij}-\mu^*_{i})^2-(\hat\mu^*_{ij}-\mu^*_{i})^2 \mid b^*_{ij}) & \text{(bottom right panel of Figure 4)}
\end{array}
$$

The two panels on the left side of Figure 4 show the frequentist perspective, where we condition on the true parameter values. We see that the Bayesian estimator performs better for small values of the parameter, while the naive estimator performs better for large values. The two panels on the right side of Figure 4 show the Bayesian perspective, where we condition on the observed b^*_{ij}. We see that from the Bayesian perspective, the Bayesian estimators always perform better.

3.5 Cross validation

We have done our best to make sure that the synthetic CDSR closely resembles the true CDSR, so that the results of the previous sections also apply to the true CDSR. However, we cannot exclude the possibility that there are some systematic differences between the synthetic and true datasets which affect the relative performance of the Bayesian and naive estimators. However, we can conduct some additional checks that do not require synthetic CDSR.

First, since the estimates b_{ij} are unbiased for the β_{ij} the RMSE is approximately $\sqrt{\frac{1}{18368}\sum_{ij} s^2_{ij}} = 0.38$, which agrees with the value of 0.38 from Table 2. The b_{ij} are also unbiased for the μ_i, so the RMSE can be estimated by $\sqrt{\frac{1}{18368}\sum_{ij} s^2_{ij} + \tau^2_{i,\text{mle}}} = 0.50$, which agrees reasonably well with the value of 0.53 in Table 3.

Second, the first column in Table 3 provides an approximation of the difference in mean squared errors across the CDSR,

$$\frac{1}{N}\sum_{ij}(b_{ij}-\mu_i)^2-(\hat\beta_{ij}-\mu_i)^2, \tag{10}$$

by the difference across the synthetic CDSR,

$$\frac{1}{N}\sum_{ij}(b^*_{ij}-\hat\mu_i)^2-(\hat\beta^*_{ij}-\hat\mu_i)^2 = 0.28-0.12=0.16. \tag{11}$$

Despite μ_i not being observed, there is an alternative, direct way to estimate (10). We start by constructing a third estimator of μ_i which is unbiased and independent of both b_{ij} and $\hat{\mu}_{ij}$. We leave out study j from meta-analysis i and run a simple fixed effects meta-analysis on the remaining $n_i - 1$ studies to obtain an estimate $\hat{\mu}_i^{-j}$ of the population average effect μ_i. This estimator is unbiased for μ_i because the individual study estimates are. Moreover, it is independent of both b_{ij} and $\hat{\mu}_{ij}$ because it is based on different studies. In [12], we proved the following proposition

Proposition 1. *Consider three estimators T_0, T_1 and T_2 of a parameter θ. Suppose that, conditionally on θ, T_0 is unbiased and independent of T_1 and T_2. Then*

$$\mathbb{E}(T_1 - T_0)^2 - \mathbb{E}(T_2 - T_0)^2 = \mathbb{E}(T_1 - \theta)^2 - \mathbb{E}(T_2 - \theta)^2, \tag{12}$$

where the expectations are with respect to arbitrary distributions of T_0, T_1, T_2, and θ (as long as the expectations are well-defined and finite).

Now, if we take $T_0 = \hat{\mu}_i^{-j}$, $T_1 = b_{ij}$, $T_2 = \hat{\beta}_{ij}$ and $\theta = \mu_i$, then it follows that (as N approaches infinity)

$$\frac{1}{N}\sum_{ij}(b_{ij} - \mu_i)^2 - (\hat{\beta}_{ij} - \mu_i)^2 \approx \frac{1}{N}\sum_{ij}(b_{ij} - \hat{\mu}_i^{-j})^2 - (\hat{\beta}_{ij} - \hat{\mu}_i^{-j})^2. \tag{13}$$

We can compute the right side directly from the original CDSR. We find that it equals 0.19. This is reasonably close to the difference of 0.16 from Table 3. This strengthens our confidence in the results from the synthetic CDSR.

4 Discussion

4.1 Using these results in applied research

Between-study variation of the treatment effect is often present in systematic reviews. Such heterogeneity may be due to differences in study populations, methodologies, or measurement techniques. In a random effects meta-analysis the uncertainty due to between-study variation can be accounted for, but it remains hidden when we have only a single trial. We propose a Bayesian approach, which we refer to as a "meta-analysis of a single trial," where we estimate the distribution of treatment effects and heterogeneity across 1,635 meta-analyses from the Cochrane Database of Systematic Reviews (CDSR). Taking these estimated distributions as prior information provides a substantial improvement in performance both for estimating the effect in the trial (Table 2) and for estimating the population average effect among similar trials (Table 3). The Bayesian meta-analysis of a single trial can easily be done in `R` by using package `baggr` [13].

The Bayesian approach results in a large reduction of the root mean squared error across the trials of the CDSR compared to the usual unbiased estimator. This is to be expected for shrinkage estimation, and we have previously obtained similar results [12]. Here, we want to draw special attention to the substantial lack of coverage of the usual confidence interval for the population average effect; see Table 3. This is due to failure to account for the heterogeneity. In contrast, the coverage of the Bayesian uncertainty interval is equal to its nominal level.

Figure 1 shows that a single trial essentially never provides certainty about the sign of population average effect. This is a strong argument for the need for replication studies.

Since our prior distributions refer to the population of trials in the CDSR, our posterior statements can be interpreted in terms of random sampling from the CDSR. So, for example, we can say that if we randomly select a trial from the CDSR (or from the population of all the trials that could be in the CDSR) and we observe that the estimated treatment effect is $b = 0.7$ with standard error 0.7, then the probability that the true effect in that trial is also positive is 72% (cf section 2.2)

About 75% of the meta-analyses in the CDSR have 5 or fewer studies [4]. When a meta-analysis consists of so few studies, it is clear that the heterogeneity cannot be estimated reliably without additional information [7]. Similarly as in the case of a single study that we outlined here, we should expect that doing a Bayesian meta-analysis with informative priors will improve inference. We intend to further evaluate the performance of this approach in a separate study.

4.2 Bayesian meta-analysis

Statistical practice—Bayesian and otherwise—has incoherence with respect to the number of studies K in a meta-analysis or, more generally, the number of groups in a multilevel model. We can see this by starting with a large K and then seeing what happens as it decreases.

When K is large, say larger than 10, the study-level variance and thus the optimal shrinkage factor can be well estimated from the data, or if the meta-analysis includes individual and study-level predictors, the unexplained group-level variance can be well estimated.

As K becomes smaller, there is more uncertainty in the study-level variance parameter, and prior information on that parameter becomes more relevant to determining the amount of shrinkage. When K is between 5 and 10, the prior on τ can make more of a difference: first by providing area-specific prior information and second through the regularization properties of weakly informative priors that (probabilistically) constrain the low and high ends of the distribution. A regularizing prior on the high end can be necessary to reduce the upper tail of the posterior for τ; in full Bayesian inference with a flat prior, the resulting long tail manifests itself by giving some probability of essentially no shrinkage, leading to wide uncertainties for the effects in individual studies. If τ is estimated using a marginal posterior mode, it can also be helpful to use zero-avoiding priors [2] as otherwise the point estimate for τ can be very noisy (zero in some cases and high in others), leading to meta-analyses that uncontrollably swing between complete pooling and very little shrinkage in otherwise similar cases. Another advantage of an informative prior is that it reduces the influence of one or two outlying studies.

When K is very small, between 3 and 5, it is still possible to perform Bayesian inference on τ with a flat prior, but the resulting posterior mode is very noisy and the full Bayesian posterior for τ will have a long right tail [5], and so in practice an informative prior for the group-level variance is necessary to avoid the meta-analysis procedure yielding unreasonable results. Setting τ equal to zero to perform a so-called fixed or common effects meta-analysis amounts to using a extremely strong and unrealistic prior.

With $K = 2$, the mathematical situation changes: a flat prior on τ yields an improper posterior distribution with an infinite right tail. This is related to the result from James-Stein theory that the no-shrinkage estimate is admissible when $K < 3$. In practice, though, there is no sharp boundary between 2 and 3 studies, as in either case we want to be using a strong prior for τ. At $K = 2$ we should expect the prior for τ to dominate to the extent that the amount of shrinkage is determined much more by the prior—that is, by the population of studies being considered as the reference class—than by the observed spread in the data.

When $K = 1$, the problem doesn't look like meta-analysis at all: it is just inference from a single study. This leads to the paradox that removing information can be expected to decrease reported uncertainty. The paradox is resolved if we note that the estimand has changed. The solution is to consider questions of meta-analysis and between-study variation even in a single study. In other words, this means placing the problem in a hierarchical context: even when multiple data sources are not available, we can still include a prior on between-study variation. This was already going to be necessary with $K = 2$ or 3, so why not do this with $K = 1$ also?

Finally, $K = 0$ corresponds to the setting where no studies are available on a problem of interest, so that the posterior is determined entirely by the prior. This can be viewed as a sort of thought experiment, representing the information being assumed from nothing but the general class of problems under study.

5 Reproducibility

The results in this paper are fully reproducible with the `R` code provided in GitHub repository at `github.com/wwiecek/singletrial`. The data are publicly available at `https://osf.io/xjv9g/`. Below we provide the code snippet for calculating probability of positive effects in a trial or population; see the example in section 2.2.

```
library(baggr)
set.seed(123)
bg <- baggr(data.frame(tau = 0.7, se = 0.7),
            prior_hypermean = student_t(5.18, 0, 0.37),
            prior_hypersd = lognormal(-1.82, 0.90))
print(bg)
mean(study_effects(bg)[,1,1] > 0)     # positive effect in the trial?
mean(hypermean(bg,summary=FALSE) > 0) # positive pooled effect?
```

6 Acknowledgement

We thank Simon Schwab and Rebecca Turner for their comments on an early draft of the paper. Witold Więcek received funding from FTX Future Fund regranting program to pursue research into meta-analytic methods. Andrew Gelman's work was supported by ONR grant N000142212648.

7 Declaration of conflicting interests

The authors declare that there are no conflicts of interest.

References

[1] Chittaranjan Andrade. Mean difference, standardized mean difference (smd), and their use in meta-analysis: as simple as it gets. *The Journal of clinical psychiatry*, 81(5):11349, 2020.

[2] Yeojin Chung, Sophia Rabe-Hesketh, Andrew Gelman, Jingchen Liu, and Vincent Dorie. A nondegenerate estimator for hierarchical variance parameters via penalized likelihood estimation. *Psychometrika*, 78:685–709, 2013.

[3] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 2013.

[4] Jonathan Davey, Rebecca M Turner, Mike J Clarke, and Julian Higgins. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. *BMC Medical Research Methodology*, 11(1):1–11, 2011.

[5] Andrew Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1:515–533, 2006.

[6] Julian PT Higgins and Anne Whitehead. Borrowing strength from external trials in a meta-analysis. *Statistics in medicine*, 15(24):2733–2749, 1996.

[7] Evangelos Kontopantelis, David A Springate, and David Reeves. A re-analysis of the cochrane library data: the dangers of unobserved heterogeneity in meta-analyses. *PloS one*, 8(7):e69930, 2013.

[8] Eleanor M Pullenayegum. An informed reference prior for between-study heterogeneity in meta-analyses of binary outcomes. *Statistics in Medicine*, 30(26):3082–3094, 2011.

[9] Simon Schwab. Re-estimating 400,000 treatment effects from intervention studies in the Cochrane Database of Systematic Reviews [data set]. *Open Science Framework*, 2020. `https://doi.org/10.17605/OSF.IO/XJV9G`.

[10] Stan Development Team. Stan modeling language user's guide. 2023. `https://doi.org/10.17605/OSF.IO/XJV9G`.

[11] Rebecca M Turner, Jonathan Davey, Mike J Clarke, Simon G Thompson, and Julian PT Higgins. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *International Journal of Epidemiology*, 41(3):818–827, 2012.

[12] Erik W van Zwet, Lu Tian, and Robert Tibshirani. Evaluating a shrinkage estimator for the treatment effect in clinical trials. *Statistics in Medicine*, 2023.

[13] Witold Więcek and Rachael Meager. *baggr: Bayesian Aggregate Treatment Effects*, 2021. R package version 0.6.9, `https://CRAN.R-project.org/package=baggr`.