Meta-analysis based on Bayesian model averaging: Principles and implementation

Authors: Ziwei Ren, Zheng Liu, Hu Chuan-Peng, Zheng Liu, Hu Chuan-Peng

Date: 2026-05-17T15:04:47+00:00

Abstract

Meta-analysis, as an important statistical method for synthesizing existing research findings, is widely applied in quantitative research. However, researchers often face dilemmas in model selection during the analysis process: when dealing with between-study heterogeneity, a choice must be made between fixed-effect and random-effects models; when addressing potential publication bias, one must select from various correction models. Currently, there is still a lack of unified standards for these model selections. Bayesian Model Averaging (BMA) within the Bayesian statistical framework provides a new perspective for solving these issues: it incorporates different statistical models into the same model space, quantifies the degree of data support for each model, and weights the effect sizes accordingly, thereby avoiding the uncertainty inherent in single-model selection. BMA-based meta-analysis can simultaneously test three key hypotheses (the existence of an effect, the existence of heterogeneity, and the existence of publication bias) and achieve robust estimation of effect sizes through model averaging. This method can be implemented via the open-source software JASP or the R language, providing researchers with a new alternative for conducting meta-analyses.

Full Text

Preamble

Meta-Analysis Based on Bayesian Model Averaging: Principles and Implementation

Ziwei Ren [1], Zheng Liu [2*], Chuan-Peng Hu [1*]

(1 School of Psychology, Nanjing Normal University, Nanjing, 210097)

(2 School of Humanities and Social Science, The Chinese University of Hong Kong (Shenzhen), Shenzhen, 518172)

Abstract

Meta-analysis, as an essential statistical method for synthesizing existing research findings, is widely applied in quantitative research. However, during the analysis process, researchers often encounter challenges such as data heterogeneity and publication bias, which can affect the robustness of the conclusions. By integrating effect sizes from multiple independent studies, meta-analysis provides a more comprehensive understanding of specific scientific questions than individual studies alone.

[Figure 1: see original paper]

In recent years, the rapid development of machine learning and deep learning has introduced new perspectives and tools to the field of meta-analysis. These advanced computational techniques allow for more sophisticated modeling of complex interactions and non-linear relationships that traditional linear models might overlook. Furthermore, the automation of systematic reviews through natural language processing (NLP) has significantly reduced the manual labor required for screening literature and extracting data.

Despite these advancements, the quality of a meta-analysis remains fundamentally dependent on the quality of the primary studies included. Issues such as the "file drawer problem," where non-significant results are less likely to be published, continue to pose a threat to the validity of synthesized results. Therefore, it is crucial for researchers to employ rigorous sensitivity analyses and bias detection methods, such as funnel plots and Egger' s regression, to ensure the integrity of their findings. As the field evolves, the integration of more transparent reporting standards and open-science practices will be vital for the continued advancement of meta-analytic methodology.

Researchers frequently face a dilemma regarding model selection: when addressing heterogeneity between studies, they must choose between fixed-effects and random-effects models; furthermore, when dealing with potential publication bias, they must decide on appropriate adjustment methods. Currently, there remains a lack of unified standards for this model selection process. Within the framework of Bayesian statistics, Bayesian Model Averaging (BMA) provides a robust approach to address this challenge.

BMA provides a novel approach by incorporating various statistical models into a unified model space and quantifying the uncertainty associated with each individual model. This methodology allows for a more robust synthesis of information, ensuring that the final estimates account for the inherent variability and potential misspecification across different modeling frameworks. The degree of data support is used to weight effect sizes, thereby avoiding the uncertainty inherent in selecting a single model. Meta-analysis based on BMA can simultaneously account for model uncertainty and provide a more robust estimation of the overall effect.

We test three key hypotheses—the existence of an effect, the presence of het-

erogeneity, and the existence of publication bias—and implement a robust estimation of the effect size using a model-averaging approach. This method can be implemented through the open-source software JASP or the R programming language, providing researchers with a new alternative for conducting meta-analyses.

1.1 Testing Key Hypotheses

The first objective of this analysis is to rigorously evaluate the empirical evidence regarding the phenomenon under study. We focus on three fundamental questions that often challenge the validity of meta-analytic results. First, we examine whether a statistically significant effect exists across the body of literature. Second, we test for the presence of heterogeneity, determining whether the observed variation in effect sizes exceeds what would be expected by sampling error alone. Third, we assess the presence of publication bias, which occurs when the likelihood of a study being published depends on the statistical significance or direction of its results. By addressing these three dimensions, we provide a comprehensive diagnostic of the current state of the research field.

1.2 Robust Estimation via Model Averaging

To ensure the reliability of our findings, we employ a model-averaging framework to achieve a robust estimation of the overall effect size. Traditional meta-analysis often relies on a single "best-fitting" model, which may ignore the uncertainty inherent in model selection. In contrast, our approach integrates results across multiple candidate models—including those that account for different levels of heterogeneity and various mechanisms of publication bias. By weighting the estimates from these models based on their relative fit to the data, we produce a posterior distribution for the effect size that is less sensitive to the specific assumptions of any individual model. This methodology provides a more stable and credible synthesis of the evidence, particularly in the presence of conflicting study results or potential reporting biases.

1 Introduction

Meta-analysis achieves a quantitative synthesis of findings by statistically integrating effect sizes from multiple independent studies. This methodology provides a more comprehensive and objective conclusion than individual studies alone, effectively increasing statistical power and resolving inconsistencies across the existing literature. A meta-analysis is a periodic and quantitative summary of research progress on a specific topic [**?**, **?**, **?**, **?**].

Results from different studies are transformed into a unified effect size metric (such as standardized mean difference, correlation coefficients, etc.) to evaluate the overall magnitude of the effect. By aggregating data across various research contexts, this approach increases statistical power and improves the precision of effect estimates. It allows researchers to identify patterns, resolve discrepancies

between conflicting findings, and explore potential moderators that influence the strength or direction of the observed phenomena. Since Glass first proposed the concept in 1976, meta-analysis has served as a critical method for accumulating evidence in fields such as psychology, education, and medicine.

Meta-analysis is essentially a hierarchical model, in which the various studies included are treated as different units within the same model. This hierarchical structure allows researchers to account for both within-study and between-study variability. Ideally, the lowest level of the model should consist of observational data from individual subjects. However, because original literature typically only reports results at the study level, we must account for this limitation. When individual-level data are unavailable, researchers rely on aggregated statistics, necessitating the use of meta-analytic techniques within a hierarchical framework to synthesize findings while accounting for within-study and between-study variability.

The methodology relies on summary data (such as effect sizes and their corresponding standard errors) rather than individual participant data. By weighting individual studies—typically by the inverse of their variance—meta-analysis provides a more precise estimate of the true effect size than any single study could achieve alone. These models utilize standard errors to quantify within-study variation resulting from sampling randomness. [Figure 1: see original paper] illustrates the most typical structure of a two-layer meta-analysis.

In practical applications, when a single study reports multiple effect sizes or other forms of dependency structures exist, the model can be further extended into a multilevel framework (Van den Noortgate et al., 2013, 2015). This "data nesting" approach allows for a more nuanced estimation of effect sizes by partitioning variance across different levels. This structure of "research-on-research" can be implemented within both the classical frequentist framework and the Bayesian framework.

Note: This example model incorporates two levels of variation: (1) Inter-study variation (Level 2): The overall distribution of effect sizes is denoted as $N(\mu, \tau^2)$, where μ is the overall average effect and τ^2 represents the between-study variance. (2) Intra-study variation (Level 1): The individual study distributions represent the sampling error within each study, typically denoted as $N(\theta_i, \sigma_i^2)$, where θ_i is the true effect size for study i and σ_i^2 is the within-study variance.

[Figure 2: see original paper]

Within-study variation (Level 1): The bottom layer represents individual subjects within each study, where individual data are aggregated into the observed summary effect size y_i. The width of the sampling distribution curves reflects the within-study variance σ_i^2. Given the true effect size θ_i, the sampling distribution of the observed value y_i is defined as $y_i \sim N(\theta_i, \sigma_i^2)$.

Between-study variation (Level 2): The top layer represents the distribution of true effect sizes across different studies. The distribution of the true effect θ_i is

characterized by the overall mean μ and the between-study variance τ^2. Assuming the absence of publication bias, the observed effect sizes y_i are distributed normally around the population mean effect μ.

Meta-analysis can be conducted within both the traditional frequentist framework and the Bayesian framework. From a frequentist perspective, the overall effect size is treated as a fixed but unknown true parameter [**?**, **?**]. This framework has limitations when dealing with the hierarchical structure: it estimates the between-study heterogeneity τ as a fixed constant, which fails to account for the uncertainty inherent in the estimation process. This may lead to overly optimistic confidence intervals and an increased risk of Type I errors.

In contrast, the Bayesian framework treats both the overall effect size μ and the heterogeneity τ as random variables represented through distributions. The parameters governing these distributions are referred to as population parameters or hyperparameters. This fully probabilistic framework allows the uncertainty inherent in heterogeneity estimation to be preserved within the estimation of effect sizes (Grant & Di Tanna, 2025; Harrer et al., 2021; Kruschke & Liddell, 2018).

When conducting a meta-analysis, researchers face two primary challenges: heterogeneity and publication bias. Heterogeneity refers to the true differences in effect sizes across different studies (Sedgwick, 2015). Publication bias occurs when results that are statistically significant are preferentially published, leading to an overestimation of effect sizes [**?**, **?**]. Researchers must choose between fixed-effect models ($\tau^2 = 0$) and random-effects models ($\tau^2 > 0$), and select from various correction methods such as PET-PEESE, publication selection models, or the trim-and-fill method. However, these models lack unified selection criteria, and different approaches may yield inconsistent conclusions (Almalik, 2025; Carter et al., 2019; Ioannidis, 2008).

Bayesian Model Averaging (BMA) addresses these issues by incorporating all reasonable candidate models into a unified model space. The posterior probability of each model is calculated and used as a weight to determine the overall effect. This directly integrates model selection uncertainty into statistical inference (Berkhout et al., 2024; Fragoso et al., 2018; Hinne et al., 2020). Robust Bayesian Meta-Analysis (RoBMA) extends this by simultaneously accounting for heterogeneity and publication bias, and is available through JASP and R packages [**?**, **?**, **?**, **?**].

2 Foundations of Bayesian Meta-Analysis

The fundamental difference between Bayesian and frequentist inference lies in their definitions of probability. Frequentism views probability as the long-term limit of an event' s relative frequency, treating parameters as fixed constants. Bayesian inference interprets probability as a measure of the degree of belief or uncertainty. Parameters are treated as random variables characterized by probability distributions. By combining a prior distribution $p(\theta)$ with a likelihood

function $p(\text{data}|\theta)$, Bayesians derive a posterior distribution $p(\theta|\text{data})$:

$$p(\theta|\text{data}) = \frac{p(\text{data}|\theta)p(\theta)}{p(\text{data})}$$

This process allows for iterative learning as new data becomes available (van de Schoot et al., 2014, 2021). Bayesian inference ensures research conclusions are not limited to binary hypothesis testing but provide direct probabilistic interpretations.

In hierarchical models, individual-level parameters are assumed to follow a higher-level distribution governed by hyperparameters (population mean μ and heterogeneity τ). The joint posterior distribution (omitting the normalization constant) is:

$$p(\theta, \mu, \tau|\text{data}) \propto p(\text{data}|\theta)p(\theta|\mu, \tau)p(\mu, \tau)$$

Here, $p(\theta|\mu, \tau)$ is the hierarchical prior, and $p(\mu, \tau)$ is the hyperprior. This structure allows for "shrinkage," where individual estimates are pulled toward the population mean, improving reliability (Lee, 2011; Veenman et al., 2024). Since these models often lack closed-form solutions, researchers rely on Markov Chain Monte Carlo (MCMC) methods for approximation (Hu Chuan-Peng & Wang Ji-Xian, 2025).

3 Model Selection Dilemmas and BMA Solutions

Researchers face two primary dilemmas: choosing between fixed-effect and random-effects models, and selecting a publication bias correction method. Traditional procedures involve selecting a single path, which ignores model uncertainty. BMA avoids this by weighting all candidate models based on their posterior model probabilities.

3.1 Addressing Heterogeneity with BMA

BMA constructs a model space based on the existence of an effect (μ) and the existence of heterogeneity (τ). This results in four models (Gronau et al., 2017, 2021):

1. **Model** $M_1(H_0^f)$**:** Fixed-effects null model ($\mu = 0, \tau = 0$).
2. **Model** $M_2(H_1^f)$**:** Fixed-effects alternative model ($\mu \neq 0, \tau = 0$).
3. **Model** $M_3(H_0^r)$**:** Random-effects null model ($\mu = 0, \tau > 0$).
4. **Model** $M_4(H_1^r)$**:** Random-effects alternative model ($\mu \neq 0, \tau > 0$).

By weighting these models, BMA provides a more robust characterization of the data structure than single-model selection.

3.2 Publication Bias Correction and RoBMA

RoBMA expands the model space further by incorporating publication bias correction models (e.g., selection models based on p-values). This allows for an assessment of whether publication bias exists while simultaneously considering model uncertainty across different theoretical frameworks. The posterior distribution of the effect size Δ is:

$$P(\Delta|D) = \sum_{k=1}^{K} P(\Delta|D, M_k)P(M_k|D)$$

This multi-path analysis avoids the biases inherent in manually selecting a single analysis trajectory.

4 Implementation and Reporting

RoBMA can be implemented in JASP by assigning effect sizes and standard errors, specifying prior distributions for effect size, heterogeneity, and publication bias, and running MCMC simulations. Researchers should report the posterior probabilities for each model and the Model-Averaged results, including Bayes Factors (BF_{10} for effect, BF_{rf} for heterogeneity, and BF_{pb} for publication bias). Adhering to reporting standards like PRISMA or PRIOR-MA ensures transparency and reproducibility [**?**, **?**, **?**].

Note: Figure translations are in progress. See original paper for figures.