scientific reports

Check for updates

OPEN

Goodness-of-fit testing for meta-analysis of rare binary events

Ming Zhang[1], Olivia Y. Xiao[2], Johan Lim[3] & Xinlei Wang[1,4,5]✉

Random-effects (RE) meta-analysis is a crucial approach for combining results from multiple independent studies that exhibit heterogeneity. Recently, two frequentist goodness-of-fit (GOF) tests were proposed to assess the fit of RE model. However, they tend to perform poorly when assessing rare binary events. Under a general binomial-normal framework, we propose a novel GOF test for the meta-analysis of rare events. Our method is based on pivotal quantities that play an important role in Bayesian model assessment. It further adopts the Cauchy combination idea proposed in a 2019 JASA paper, to combine dependent p-values computed using posterior samples from Markov Chain Monte Carlo. The advantages of our method include clear conception and interpretation, incorporation of all data including double zeros without the need for artificial correction, well-controlled Type I error, and generally improved ability in detecting model misfits compared to previous GOF methods. We illustrate the proposed method via simulation and three real data applications.

Meta-analysis is a valuable technique used in various fields, including medicine, biology, social sciences, and ecology, to combine information from multiple studies to increase inference reliability. A random-effects model (REM) is a popular choice in a meta-analysis, which assumes that the actual effect sizes of component studies θ_is follow a normal distribution with an overall mean θ_0 and variance τ^2 (often referred to as the heterogeneity parameter). When $\tau^2 = 0$, a REM is reduced to a fixed-effect model (FEM) with $\theta_i \equiv \theta_0$. REMs are preferred over FEMs in most scenarios because they account for the heterogeneity among studies and are, therefore, applicable to a broader range of scenarios[1].

Among REMs, a generic model widely employed for binary and continuous outcomes uses the normal-normal hierarchical structure. For study i, let y_i be the observed effect size (for binary data, y_i is typically the log odds ratio), and σ_i^2 denotes the within-study variance (i.e., the sampling variation in study i). The generic model specifies $y_i|\theta_i, \sigma_i^2 \sim \mathrm{N}(\theta_i, \sigma_i^2)$ and $\theta_i \sim \mathrm{N}(\theta_0, \tau^2)$. However, for rare binary outcomes, the normal approximation for y_i given θ_i and σ_i^2 may not work well due to the sparsity or small sample sizes. Alternatively, the binomial-normal (BN) hierarchal structure is a popular substitute for the normal approximation. It assumes that the number of observed events in the treatment (control) group for study i, denoted by $x_{i2}(x_{i1})$, follows a binomial distribution with the total number of subjects $n_{i2}(n_{i1})$ and event probability $p_{i2}(p_{i2})$. The logit transformed probabilities are then assumed to be distributed normally in the second hierarchy, where the log odds scale measures the effect size θ_i. Several variations of the BN framework have been proposed. Bhaumik et al.[2] assumed $\mathrm{logit}(p_{i1}) = \mu_i$ and $\mathrm{logit}(p_{i2}) = \mu_i + \theta_i$, where $\mu_i \sim \mathrm{N}(\mu_0, \sigma^2)$ denotes logit-transformed background incidence rate for study i. Smith, Spiegelhalter, and Thomas[3] considered the equal variance between the control and treatment group by defining $\mathrm{logit}(p_{i1}) = \mu_i - \theta_i/2$ and $\mathrm{logit}(p_{i2}) = \mu_i + \theta_i/2$. Li and Wang[4] proposed a more flexible model by defining $\mathrm{logit}(p_{i1}) = \mu_i - \omega\theta_i$ and $\mathrm{logit}(p_{i2}) = \mu_i + (1-\omega)\theta_i$, where the new parameter ω adjusts the variance ratio between two arms and the previous two models can be viewed as special cases by assigning $\omega = 0$ and 1/2, respectively. To avoid the assumption of independency between μ_i and θ_i, Houweilingen, Zwinderman and Stijnen[5] proposed the use of a bivariate normal distribution for modeling $(\mathrm{logit}(p_{i1}), \mathrm{logit}(p_{i2}))$, which allows any correlation structure between $\mathrm{logit}(p_{i1})$ and $\mathrm{logit}(p_{i2})$ in order to test the effects of each variable.

All the models discussed above make a common assumption that the true effect sizes $\theta_i's$ follow a normal distribution $\theta_i \sim \mathrm{N}(\theta_0, \tau^2)$. While this assumption is convenient for mathematical purposes, it may not always

[1]Department of Statistics and Data Science, Southern Methodist University, Dallas, Texas 75205, USA. [2]Highland Park High School, Dallas, Texas 75205, USA. [3]Department of Statistics, Seoul National University, Seoul 08826, Korea. [4]Department of Mathematics, University of Texas at Arlington, Arlington, Texas 76019, USA. [5]Center for Data Science Research and Education, College of Science, University of Texas at Arlington, Arlington, Texas 76019, USA. ✉email: xinlei.wang@uta.edu

hold in reality as the distribution of true effect sizes across different studies could have any shape. Therefore, conducting a goodness-of-fit (GOF) test is crucial before drawing conclusions or making inferences, since a misspecified model may yield misleading results[6,7]. Researchers have come up with various solutions to test the normality of models. Recently, Chen, Zhang, and Li[8] proposed a parametric bootstrap-type GOF test, mainly focused on the generic REM. Subsequently, Wang and Lee[9] developed a standardization framework to evaluate the normality assumption. It avoids the need to generate reference distributions and is therefore computationally efficient. However, their methods require continuity corrections when encountering single or double-zero studies, which can impact both Type I error rates and statistical power. Furthermore, those who previously proposed methods did not investigate their approaches numerically under different background incidence rates, especially when dealing with rare binary outcomes. This is an interesting and important aspect to explore in meta-analyses of binary outcomes.

In terms of Bayesian alternatives, no Bayesian approach has been considered for GOF testing in meta-analysis to our knowledge. We propose a novel GOF test for meta-analysis that utilizes the pivotal quantity (PQ) methodology proposed for Bayesian model assessment[10,11], and adapts the Cauchy combination test[12], which combines dependent p values computed using posterior draws from Markov chain Monte Carlo (MCMC), to inform the final conclusion. A pivotal quantity is a function of data and model parameters whose distribution does not depend on unknown parameters. For instance, suppose $\theta = (\mu, \sigma) \sim \pi$, and $\theta_0 = (\mu_0, \sigma_0)$ is a random vector drawn from density π, which generates the normal data $\mathbf{x} = \{x_1, ..., x_n\}$. Then $f(\mathbf{x}, \mu_0, \sigma_0) = \sum_{i=1}^{n} (\frac{x_i - \mu_0}{\sigma_0})^2 \sim \chi_n^2$ is a pivotal quantity. Let $\tilde{\mu}$ and $\tilde{\sigma}$ be samples from the corresponding posterior distribution $p(\mu, \sigma | \mathbf{x})$. The PQ method is constructed based on the fact that $f(\mathbf{x}, \mu_0, \sigma_0)$ and $f(\mathbf{x}, \tilde{\mu}, \tilde{\sigma})$ are identically distributed; that is, $f(\mathbf{x}, \tilde{\mu}, \tilde{\sigma}) \sim \chi_n^2$.

Our proposed method, called Improved Pivotal Quantities (IPQ), can detect a model failure at all levels in hierarchical models without extra computational cost. Additionally, it can be easily incorporated into standard Bayesian implementations and automatically accounts for all available data without requiring artificial corrections for rare binary events. While our method is suitable for general purposes, we focus primarily on its application for meta-analyses of rare binary outcomes.

The rest of this article is organized as follows. We first review Bayesian techniques that assess model adequacy, followed with a brief introduction to pivotal quantities and the Cauchy combination test. We then introduce our proposed method based on the generalized REM in Houweilingen, Zwinderman and Stijnen[5] for meta-analysis of binary events. We also describe the Bayesian implementation of our method, including adapting the proposed IPQ method within the MCMC algorithm and considering different bivariate covariance priors. In the simulation section, we conduct simulation studies to evaluate our method's performance in terms of Type I error rates and statistical power and compare it with other existing GOF methods. We also evaluate four different covariance priors based on estimating the overall treatment effect, the inter-study heterogeneity, and the correlation coefficient. In data examples section, we illustrate our method using three real data examples. The first example utilizes handedness and eye-dominance data from 54 studies, the second one employs Type 2 diabetes mellitus and gestational diabetes data from 20 studies, and the third uses GSTP1 gene and lung cancer data from 44 studies. We then end this paper with conclusions and discussions.

Review of related bayesian work

In current practice, Bayesian model diagnostics mainly fall into three categories: prior predictive, posterior predictive, and pivotal quantity-based approaches. See Figure 1 for illustration.

Prior and posterior predictive checks

Suppose $\mathbf{x}$ has a distribution function specified by $p(\mathbf{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents the parameters of a model (say $\mathscr{M}$) under study. Let $\mathbf{x}^{\text{obs}}$ denote the observed data and $\mathbf{x}^{\text{rep}}$ denote the replicated data that are generated to mimic real data. Box[13] recommended using the prior predictive distribution, $p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$, as a reference distribution to generate $\mathbf{x}^{\text{rep}}$ for comparing with $\mathbf{x}^{\text{obs}}$. The steps to obtain the prior predictive distribution are illustrated in Figure 1A. Given $\boldsymbol{\theta}^{\text{rep}_j}$ drawn from the prior $p(\boldsymbol{\theta})$ for $j = 1, .., R$, we draw $\mathbf{x}^{\text{rep}_j}$ from the sampling distribution $p\left(\mathbf{x}|\boldsymbol{\theta}^{\text{rep}_j}\right)$. We then use $T(\mathbf{x}, \boldsymbol{\theta})$ or $T(\mathbf{x})$, a function of data and model parameters or a function of data alone, to measure the discrepancy between data and model assumptions. Here, we take $T(\mathbf{x})$ as an example for simplicity, evaluated at both $\mathbf{x}^{\text{obs}}$ and $\mathbf{x}^{\text{rep}_j}$ for all j. The model misfit can be concluded if $T\left(\mathbf{x}^{\text{obs}}\right)$ is unlikely from the reference distribution formed by $T\left(\mathbf{x}^{\text{rep}_j}\right)$'s. However, prior predictive checks might be problematic when using improper or weakly-informative priors, which are commonly used in practice[14].

Gelman, Meng, and Stern[15] proposed model assessment using the posterior predictive distribution, defined as $p\left(\mathbf{x}^{\text{rep}}|\mathbf{x}^{\text{obs}}\right) = \int p(\mathbf{x}^{\text{rep}}|\boldsymbol{\theta}) p\left(\boldsymbol{\theta}|\mathbf{x}^{\text{obs}}\right) d\boldsymbol{\theta}$. As shown in Figure 1B, replicated data $\mathbf{x}^{\text{rep}_j}$ are generated using $\boldsymbol{\theta}^{\text{rep}_j}$ from the posterior distribution $p\left(\boldsymbol{\theta}|\mathbf{x}^{\text{obs}}\right)$. Then, the reference distribution based on the chosen discrepancy function $T(\mathbf{x})$ can be computed. The Bayesian posterior predictive p value can be obtained as $P\left[T(\mathbf{x}^{\text{rep}}) \geq T\left(\mathbf{x}^{\text{obs}}\right)|\mathbf{x}^{\text{obs}}\right]$ to quantitatively detect the model misfit.

The posterior predictive check has gained increasing popularity in Bayesian model checking due to its straightforward implementation via Monte Carlo Markov Chain (MCMC) algorithms. However, there are two major limitations associated with this type of approach. Firstly, unlike the traditional p value, the posterior predictive p value does not follow a uniform distribution under the null hypothesis of no lack of fit, making it difficult to interpret and assess the level of evidence against the null hypothesis[16]. Secondly, the method has almost no power to detect failures from the second or deeper layers in hierarchical models[11,17].

To overcome the non-uniformity problem, a potential solution is to calibrate the posterior predictive p value so that the calibrated p value follows a uniform distribution asymptotically[18]. However, the statistical power of

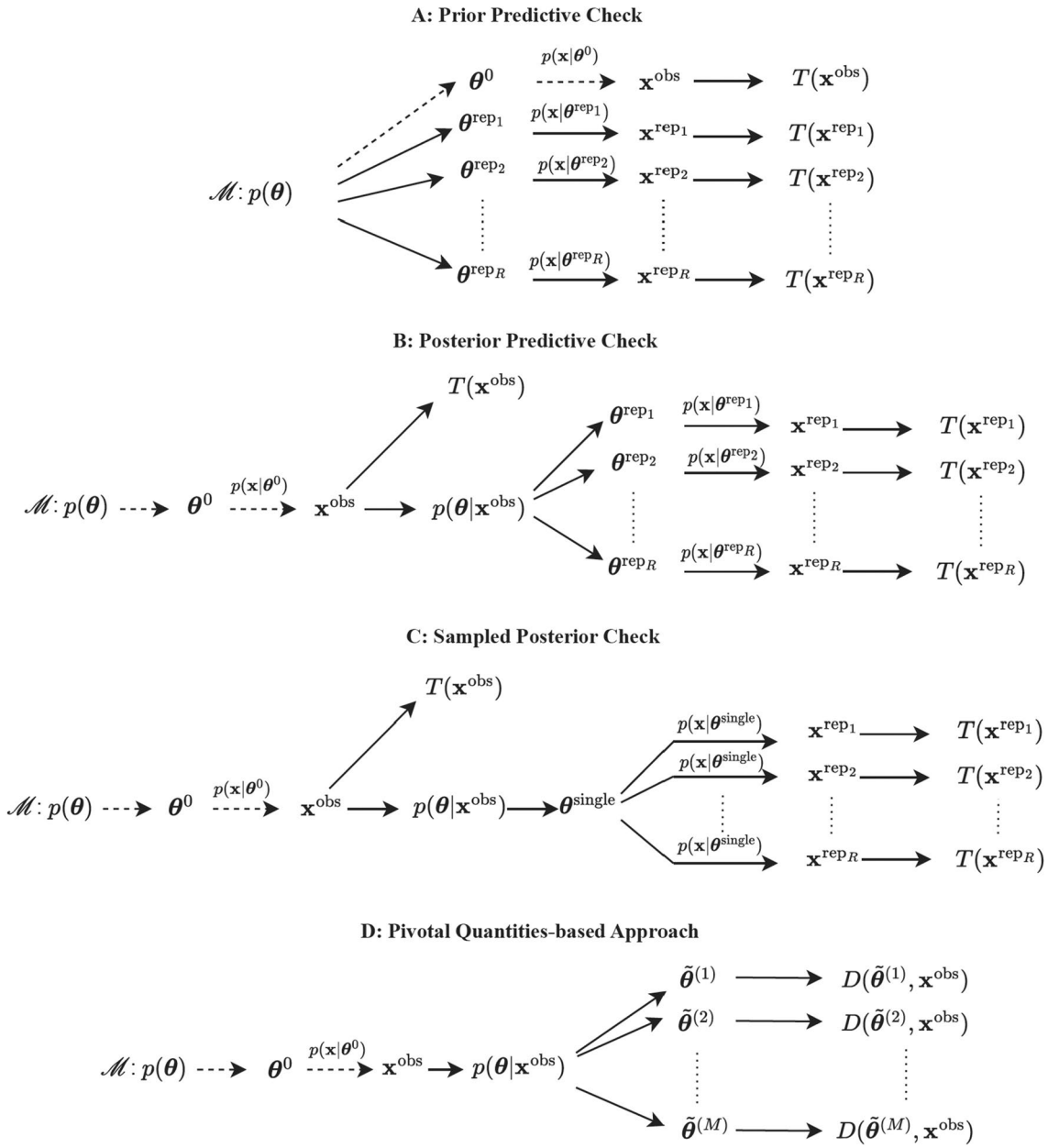

Figure 1. Schematic diagrams of different model diagnostic methods.

using the calibrated p value has not been investigated yet. To avoid this issue, Bayarri and Berger[19] proposed two new types of p values: the conditional predictive p values and the partial posterior predictive p values. Bayarri and Castellanos[16] further extended the partial posterior predictive method to test the second layer of hierarchical models, which avoids "using data twice." However, as mentioned in Johnson[20], the partial posterior strategy is typically not straightforward to implement beyond normal-family problems. More recently, Gosselin[21] and Zhang[22] recommended randomly drawing a single value $\boldsymbol{\theta}^{\text{single}}$ from the posterior distribution $\pi\left(\boldsymbol{\theta}|\mathbf{x}^{\text{obs}}\right)$ to generate $\mathbf{x}^{\text{rep}}$, namely sampled posterior check (Figure 1C). The corresponding p values is distributed uniformly when the data model is correctly specified, and the approach achieved higher power than the original posterior predictive check for detecting model misfit[21].

Pivotal quantity methodology

Johnson[10] pioneered the use of pivotal quantities (PQ) to detect model misfit, and Yuan and Johnson[11] extended upon the methodology so that it can be applied to any level of hierarchical models. Since it does not involve replicated data, there is no need to distinguish $\mathbf{x}^{\text{obs}}$ and $\mathbf{x}^{\text{rep}}$, and $\mathbf{x}$ is directly used for observed data.

A pivotal quantity, denoted by $D(\mathbf{x};\boldsymbol{\theta})$, is a function of both data $\mathbf{x}$ and model parameters $\boldsymbol{\theta}$. It possesses a sampling distribution F that is both known and invariant when evaluated at $\boldsymbol{\theta}^0$, the "true" (data-generating) value of $\boldsymbol{\theta}$; that is, $D\left(\mathbf{x};\boldsymbol{\theta}^0\right) \sim F$. Johnson[10] shows that $D\left(\mathbf{x};\tilde{\boldsymbol{\theta}}\right)$ and $D\left(\mathbf{x};\boldsymbol{\theta}^0\right)$ are identically distributed, where $\tilde{\boldsymbol{\theta}}$ is

drawn from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$. Based on this result, the approach to model assessment involves two main steps[10]. The first is to select a pivotal discrepancy measure $D(\mathbf{x};\boldsymbol{\theta})$ with a known reference distribution F, and, the second step is to evaluate the model fit by determining whether $D\left(\mathbf{x};\tilde{\boldsymbol{\theta}}\right)$ can be considered as a draw from F. However, when conducting a GOF test for the second or deeper layers in hierarchical models, one may encounter difficulties since $D(\mathbf{x};\boldsymbol{\theta})$ depends on $\mathbf{x}$, but these layers usually involve no data. For this reason, Yuan and Johnson[11] extended the method by defining the pivotal quantity D as a function of model parameters only and further showed that $D\left(\boldsymbol{\theta}^0\right)$ and $D\left(\tilde{\boldsymbol{\theta}}\right)$ have identical distributions. This allows the application of pivotal quantities and the corresponding reference distributions to diagnose model inadequacy at any level of a hierarchical model.

As shown in Figure 1D, after drawing $\tilde{\boldsymbol{\theta}}^{(i)}$ from $p(\boldsymbol{\theta}|\mathbf{x})$, $D(\mathbf{x};\boldsymbol{\theta})$ is evaluated at $\tilde{\boldsymbol{\theta}}^{(i)}$ for $i = 1, ..., M$. Then, each $D\left(\mathbf{x},\tilde{\boldsymbol{\theta}}^{(i)}\right)$ has the same distribution as $D\left(\mathbf{x},\boldsymbol{\theta}^0\right)$. For example, suppose $D\left(\mathbf{x},\boldsymbol{\theta}^0\right) \sim \mathrm{N}(0,1)$, then under the null hypothesis of no lack of fit, $D\left(\mathbf{x},\tilde{\boldsymbol{\theta}}^{(i)}\right) \sim \mathrm{N}(0,1)$ for each i marginally. To test normality, Johnson[10] suggested using a formal approach such as the Shapiro-Wilks test, and different p values from the tests on $\left(\mathbf{x},\tilde{\boldsymbol{\theta}}^{(i)}\right)$ are calculated for $i = 1, ...M$. However, combining those p values is not as straightforward as using Fisher's combination test. This is because the p values are derived from posterior samples using the same dataset and so are dependent with an unknown covariance structure.

To address this issue, Johnson[10] suggested that one could avoid generating multiple draws from the same dataset by utilizing the prior-predictive distribution from Dey et al.[17], which suggested generating 1000 replicated datasets, $\mathbf{x}^{\mathrm{rep}_i}$ for $i = 1, ..., 1000$, from $p(\mathbf{x})$ as illustrated in Figure 1A. For each replicated dataset $\mathbf{x}^{\mathrm{rep}_i}$, Bayesian data analysis is performed to obtain the corresponding posterior distribution $p(\boldsymbol{\theta}|\mathbf{x}^{\mathrm{rep}_i})$, then a single $\tilde{\boldsymbol{\theta}}^{\mathrm{rep}_i}$ is randomly sampled from the posterior, and one p values from testing the normality using the pivotal quantity is computed. This results in 1000 independent p values. Standard approaches, such as Fisher's test, can then be employed to draw a conclusion. However, this method may suffer from two limitations. Firstly, using 1000 replicated datasets can be computationally intensive since the same MCMC procedure needs to run 1000 times to draw independent posterior samples. Secondly, non-informative priors may not necessarily generate reasonable datasets. Considering these difficulties, Johnson[10] recommended finding probabilistic bounds on dependent p values using the properties of order statistics derived from Gascuel and Caraux[23] and Rychlik[24].

Let $x_{(1)}, .., x_{(M)}$ denote order statistics from a dependent sample of random variables, where each has distribution function F, and let $F_{k:M}$ denote the distribution function for the k-th order statistic out of M. Then, the bound of $F_{k:M}$ can be written as

$$F_{k:M}(t) \geq \max\left\{0, \frac{MF(t) - k + 1}{M - k + 1}\right\}. \tag{1}$$

Let $p_1, ..., p_M$ be dependent p values for $m = 1, ..., M$. Under the null, each p values should be distributed uniformly on (0, 1), implying $F(t) = t$ in Eq. (1). Let $x_i = -p_i$, Li, Wu and Feng[25] showed that the Eq. (1) becomes

$$F_{k:M}(t) \leq \min\left(1, t\frac{M}{k}\right),$$

which means that a p value upper bound for the observed k-th order statistic $p^{\mathrm{obs}}_{(k)}$ is $\min\left(1, p^{\mathrm{obs}}_{(k)}\frac{M}{k}\right)$. To avoid choosing the value of k, they suggested reporting the minimum upper bound such that $p_{\min} = \min\left\{\min\left(1, p^{\mathrm{obs}}_{(k)}\frac{M}{k}\right)\right\}_{k=1,...,M}$. Yuan and Johnson[11] advocated using the rule-of-thumb value of 0.25 as a cutoff for declaring the model misfit in practice; that is, reject the null hypothesis $\mathscr{H}_0$ if $p_{\min} < 0.25$. However, the proposal may be liberal, and our simulation studies in the simulation section show that 0.25 is not necessarily a good choice and it is hard to select an optimal cutoff to balance the trade-off between Type I error and power.

Method

The generalized REM for meta-analysis of binary events

Suppose a meta-analysis contains I independent studies, and for the i^{th} study, let $x_{i1}(x_{i2})$ be the number of observed events in the control (treatment) group, which follows a binomial distribution with the total number of subjects $n_{i1}(n_{i2})$ and corresponding event probability $p_{i1}(p_{i2})$. Let $\phi_{i1}(\phi_{i2})$ denote the logit-transformed $p_{i1}(p_{i2})$, i.e., $\phi_{ij} \equiv \ln\left(\frac{p_{ij}}{1-p_{ij}}\right)$. Then the generalized binomial-normal REM in Houweilingen, Zwinderman and Stijnen[5] can be written as

$$\begin{gathered} x_{i1} \sim \mathrm{Bin}(n_{i1}, p_{i1}),\ x_{i2} \sim \mathrm{Bin}(n_{i2}, p_{i2}), \\ \mathrm{logit}(p_{i1}) = \phi_{i1},\ \mathrm{logit}(p_{i2}) = \phi_{i2}, \\ \begin{pmatrix}\phi_{i1}\\ \phi_{i2}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right), \end{gathered} \tag{2}$$

where (ϕ_{i1}, ϕ_{i2}) is modeled by a bivariate normal distribution with an arbitrary covariance structure. We further define the treatment effect $\theta_i = \phi_{i2} - \phi_{i1}$ for study i, which follows a univariate normal distribution with an overall mean effect $\theta_0 = \mu_2 - \mu_1$ and the heterogeneity $\tau^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$.

The generalized REM builds a strong connection to many well-established models[26]. For example, the model in Li and Wang[4] and Zhang et al.[27] is a special case of model (2), yielding

$$\begin{pmatrix} \phi_{i1} \\ \phi_{i2} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu - \omega\theta \\ \mu + (1-\omega)\theta \end{pmatrix}, \begin{pmatrix} \omega^2\tau^2 + \sigma^2 & \sigma^2 - \omega(1-\omega)\tau^2 \\ \sigma^2 - \omega(1-\omega)\tau^2 & (1-\omega)^2\tau^2 + \sigma^2 \end{pmatrix} \right),$$

where in (2), $\sigma_1^2 = \omega^2\tau^2 + \sigma^2, \sigma_2^2 = (1-\omega)^2\tau^2 + \sigma^2$, and $\rho = \frac{\sigma^2 - \omega(1-\omega)\tau^2}{\sqrt{(\omega^2\tau^2+\sigma^2)\left((1-\omega)^2\tau^2+\sigma^2\right)}}$. As mentioned in the introduction, we can let ω be 0 or 0.5, which further reduces the model to the one in Bhaumik et al.[2] or Smith, Spiegelhalter, and Thomas[3], respectively. Thus, model (2) is regarded as the most generalized binomial-normal model with fewer assumptions, so we choose it as the basis to design the GOF test for detecting non-normality of θ_i's.

Let $\boldsymbol{\Theta} = \{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \mu_1, \mu_2, \boldsymbol{\Sigma}\}$ be all parameters in (2), where $\boldsymbol{\phi}_j = \{\phi_{1j}, ..., \phi_{Ij}\}$ for $j = 1, 2$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. Let $\mathbf{X} = \{x_{i1}, x_{i2}\}_{i=1}^I$ be the data. Then the full probability model is given by

$$p(\mathbf{X}, \boldsymbol{\Theta}) = \prod_{i=1}^{I} p(x_{i1}, x_{i2}|\phi_{i1}, \phi_{i2}) p(\phi_{i1}, \phi_{i2}|\mu_1, \mu_2, \boldsymbol{\Sigma}) p(\mu_1, \mu_2, \boldsymbol{\Sigma}),$$

where

$$p(x_{i1}, x_{i2}|\phi_{i1}, \phi_{i2}) = \begin{pmatrix} n_{i1} \\ x_{i1} \end{pmatrix} \begin{pmatrix} n_{i2} \\ x_{i2} \end{pmatrix} \frac{(e^{\phi_{i1}})^{x_{i1}}}{(1+e^{\phi_{i1}})^{n_{i1}}} \frac{(e^{\phi_{i2}})^{x_{i2}}}{(1+e^{\phi_{i2}})^{n_{i2}}},$$

$p(\phi_{i1}, \phi_{i2}|\mu_1, \mu_2, \boldsymbol{\Sigma})$ is the density function of the bivariate normal distribution; $p(\mu_1, \mu_2, \boldsymbol{\Sigma})$ is the joint prior distribution on hyper-parameters introduced by the bivariate normal distribution of (ϕ_{i1}, ϕ_{i2}).

The proposed GOF test

Inspired by previous research, we propose a novel GOF test and demonstrate its applicability in the context of meta-analysis of (rare) binary events. Our approach involves defining the null hypothesis, denoted by $\mathscr{H}_0$, which assumes normality for the true effect sizes θ_is, a prevailing assumption made in meta-analysis. The alternative hypothesis, denoted by $\mathscr{H}_1$, is formulated as any departure from $\mathscr{H}_0$. In other words, we aim to detect this specific departure from the bivariate normal model assumed for the second layer of the generalized REM, where we can draw our conclusion about the presence of an overall treatment effect θ_0 and between-study heterogeneity τ^2.

Let $\boldsymbol{\Theta}_i^* = \left(\phi_{i1}^*, \phi_{i2}^*, \mu_1^*, \mu_2^*, \sigma_1^{2*}, \sigma_2^{2*}, \rho^*\right)$ be data-generating parameter values for study i, and $\tilde{\boldsymbol{\Theta}}_i^{(m)}$ be the corresponding m^{th} draw from the joint posterior distribution $p(\boldsymbol{\Theta}|\mathbf{X})$ for $m = 1, ..., M$. We define a discrepancy measure to capture the deviation from $\mathscr{H}_0$, namely

$$D\left(\boldsymbol{\Theta}_i^*\right) = \frac{\phi_{i1}^* - \phi_{i2}^* - \theta_0^*}{\tau^*} = \frac{\phi_{i1}^* - \phi_{i2}^* - \left(\mu_1^* - \mu_2^*\right)}{\sqrt{\sigma_1^{2*} + \sigma_2^{2*} - 2\rho^*\sigma_1^*\sigma_2^*}},$$

which is a pivotal quantity and follows a standard normal distribution under $\mathscr{H}_0$. Furthermore, as pointed out by one of our reviewers, this measure can be viewed as the study-specific effect in units of standard deviations, for each study and each posterior draw. Then, according to Yuan and Johnson[11],

$$D\left(\tilde{\boldsymbol{\Theta}}_i^{(m)}\right) = \frac{\tilde{\phi}_{i1}^{(m)} - \tilde{\phi}_{i2}^{(m)} - \tilde{\theta}_0^{(m)}}{\tilde{\tau}^{(m)}} = \frac{\tilde{\phi}_{i1}^{(m)} - \tilde{\phi}_{i2}^{(m)} - \left(\tilde{\mu}_2^{(m)} - \tilde{\mu}_2^{(m)}\right)}{\sqrt{\tilde{\sigma}_1^{2(m)} + \tilde{\sigma}_2^{2(m)} - 2\tilde{\rho}^{(m)}\tilde{\sigma}_1^{(m)}\tilde{\sigma}_2^{(m)}}} \tag{3}$$

has a distribution identical to $D\left(\boldsymbol{\Theta}_i^*\right)$; that is, $D\left(\tilde{\boldsymbol{\Theta}}_i^{(m)}\right) \sim \mathrm{N}(0, 1)$ for every m marginally.

Conducting a standard normality test using the pivotal quantities in (3) based on a single draw is straightforward, but this sampled posterior approach can be problematic since the vagaries of randomness can produce a sample that seems unwise. Alternatively, combining multiple MCMC draws to draw a conclusion was recommended in Johnson[10] and Yuan and Johnson[11], where probabilistic bounds of the order statistics of the p values are used to combine the dependent p values. Here, we propose to use the Cauchy combination idea[12] to combine the dependent p values.

Consider p_i as the p values obtained from the i-th statistical test, and ω_i as the corresponding nonnegative weight that sums up to 1. Liu and Xie[12] introduced the Cauchy combination test and demonstrated that, subject to certain regularity conditions, the tail of a test statistic that linearly combines individual transformed p values can be well approximated by a standard Cauchy distribution under the null hypothesis. Specifically, if there are k p values, then the test statistic is given by $T = \sum_{i=1}^k \omega_i \tan\left\{\left(0.5 - p_i\right)\pi\right\}$, where the weight ω_i is typically set to $1/k$ in the absence of any prior information. The Cauchy combination test has several salient features. Firstly, the test, by leveraging the Cauchy distribution, the test has a simple analytical formula to compute the p value. Next, unlike classical Fisher's test[28] or other common tests for combining p values, such as the minimum p value test[29], the Berk-Jones test[30] and the higher criticism test[31], the Cauchy combination test handles p values from correlated statistical tests and remains valid for arbitrary correlation structures. Finally, the test works well even if one main assumption required for the test, the bivariate normality between any two test statistics generating the p values,

is not satisfied. Thus, p_is can be from non-normal typed tests (i.e., those with test statistics that are not normally distributed), such as the Shapiro-Wilk test[32], the Cramer-von Mises test[33–35] and the Anderson-Darling test[35,36].

In summary, the proposed GOF test, namely Improved Pivotal Quantities (IPQ), can be outlined by the following steps:

Step 1: Given I independent studies, randomly sample $\tilde{\boldsymbol{\Theta}}_i^{(m)}$ from the joint posterior distribution $p(\boldsymbol{\Theta}|\mathbf{X})$ via MCMC for $i = 1, ..., I$ and $m = 1, ...M$.

Step 2: Calculate $D\left(\tilde{\boldsymbol{\Theta}}_i^{(m)}\right)$ in Eq. (3) for all i and m. For each m^{th} draw, use $\left\{D\left(\tilde{\boldsymbol{\Theta}}_i^{(m)}\right)\right\}_{i=1}^{I}$ to conduct a formal normality test (e.g. Shapiro-Wilk test) to get its p values, say $p^{(m)}$.

Step 3: Compute the test statistic $T_0 = \sum_{m=1}^{M} \frac{\tan\left\{(0.5-p^{(m)})\pi\right\}}{M}$, and calculate the corresponding p values using the formula $p^* = \frac{1}{2} - \frac{\arctan T_0}{\pi}$. Then, we will reject the $\mathscr{H}_0$ if $p^* < \alpha$ with a pre-specified significance level (e.g., $\alpha = 0.01, 0.05, 0.1$).

Bayesian implementation with different covariance priors

We now pivot the discussion to prior specification and the Bayesian implementation. We use a Hamiltonian Monte Carlo (HMC) algorithm via Stan (version 2.19.1)[37] in conjunction with R[38] to fit models with different priors discussed below. For each dataset, we run the algorithm with 5000 burn-in iterations and 5000 additional sampling iterations. The convergence of MCMC chains is detected using the Gelman-Rubin diagnostic[39].

We start with the prior choices for logit-transformed mean effects μ_1 and μ_2 for the control and treatment groups, where we consider diffuse uniform priors such that $\mu_j \sim \mathrm{U}\left(\mathrm{L}_{\mu_j}, \mathrm{U}_{\mu_j}\right)$ for $j = 1, 2$. To define the range, we get rough estimates $\hat{\mu}_{ij}$ for all I studies, $\hat{\mu}_{ij} = \ln \frac{x_{ij}+0.5}{n_{ij}-x_{ij}+0.5}$. Then, we define the lower bound $\mathrm{L}_{\mu_j} = \min_{i,j}\left\{\hat{\mu}_{ij}\right\} - c$ and upper bound $\mathrm{U}_{\mu_j} = \max_{i,j}\left\{\hat{\mu}_{ij}\right\} + c$, where we let $c = 5$ as in Bai et al.[40] so that the priors are conservative enough to contain all plausible values.

Regarding the prior for the covariance matrix $\boldsymbol{\Sigma}$, several commonly used conjugate priors are available, including the independent prior (IND) that assumes mutual independence *a priori* among the elements of $\boldsymbol{\Sigma}$[41], the inverse Wishart prior (IW)[42] and the hierarchical inverse Wishart prior (HIW)[43]. Other alternatives include the scaled inverse Wishart prior (SIW)[44] and the prior based on the separation strategy (SS)[45]. The Bayesian inference of a covariance matrix is highly sensitive to different choices of priors, and several studies have compared the performance of various priors. For example, Alvarez, Niemi and Simpson[46] compared four different priors (IW, HIW, SIW and SS) in the multivariate normal model and found that the IW prior performed the worst among all the four, especially when the true variances were small. Rúa, Mazumdar and Strawderman[41] conducted extensive simulation by comparing 38 priors, including IW, HIW, and IND, with different hyper-parameter specifications in multivariate Bayesian meta-analysis models. They found that the IW prior had overall poor performance, while the HIW prior had much more consistent performance across all scenarios examined. Akinc and Vandebroek[47] focused on the same priors used in Alvarez, Niemi and Simpson[46] and investigated Bayesian inference of the covariance matrix in mixed logit models. They suggested using different priors to check the robustness of the results but recommended avoiding the IW prior. To the best of our knowledge, the impact of different covariance priors on BN models in the context of meta-analysis of rare binary events has not been investigated. Thus, we aim to address the gap and access how these priors perform under rare binary settings. Below we briefly review four classes of priors, including IW, HIW, SS , and SIW, and their performance will be assessed in the simulation section.

Inverse Wishart prior

Due to the conjugacy property, the IW prior is often used as a default choice for covariance matrices. The density function of the IW prior $\mathrm{IW}(\nu, \mathbf{H})$ is defined as $p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{(\nu+3)}{2}} \exp\left\{-\frac{1}{2}\mathrm{trace}\left(\mathbf{H}\boldsymbol{\Sigma}^{-1}\right)\right\}$, where $\nu > 0$ is the number of degrees of freedom and $\mathbf{H}$ is a symmetric scale matrix with two dimensions. The marginal distribution of the correlation parameter ρ in $\boldsymbol{\Sigma}$ is $p(\rho) \propto \left(1-\rho^2\right)^{\frac{(\nu-3)}{2}}$ when $\mathbf{H}$ is a diagonal matrix. If $\nu = 3$, ρ follows $\mathrm{U}(-1, 1)$[45]. For our model, the conditional posterior distribution of $\boldsymbol{\Sigma}$ is given by $p\left(\boldsymbol{\Sigma}|\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \mu_1, \mu_2, \mathbf{x}\right) \sim \mathrm{IW}\left(\nu + I, \mathbf{H} + \boldsymbol{\Lambda}_\mu\right)$, where

$$\boldsymbol{\Lambda}_\mu = \begin{pmatrix} \sum_{i=1}^{I}(\phi_{i1}-\mu_1)^2 & \sum_{i=1}^{I}(\phi_{i1}-\mu_1)(\phi_{i2}-\mu_2) \\ \sum_{i=1}^{I}(\phi_{i1}-\mu_1)(\phi_{i2}-\mu_2) & \sum_{i=1}^{I}(\phi_{i2}-\mu_2)^2 \end{pmatrix}.$$

While the IW prior is a popular choice in Bayesian analysis due to its mathematical convenience, it also has limitations. One issue is that selecting the appropriate degrees of freedom ν and scaled matrix $\mathbf{H}$ can be challenging. Although these are often set to default values of 3 and an identity matrix, respectively, recent studies by Rúa et al.[41] and Akinc and Vandebroek[47] have shown that these choices may not always be suitable. Another limitation is that the IW prior implies a strong relationship between variance and correlation, which can bias inference. Specifically, smaller variances are associated with correlation coefficients ρ around 0, while larger variances correspond to ρ approaching −1 or 1. This dependency can be problematic when interpreting results and drawing conclusions from statistical analyses.

Hierarchical inverse Wishart prior

Huang and Wand[43] proposed a two-layer hierarchical prior that builds upon the work of Wand et al.[48] and Armagan, Artin et al.[49], who showed that a half-t distribution can be expressed as a scale mixture of an inverse gamma distribution. In our case, the dimension of $\mathbf{\Sigma}$ is two, so that their hierarchical prior is defined as $p(\mathbf{\Sigma}|a_1, a_2) \sim \text{IW}(\nu + 1, \mathbf{H}^*)$, where $\mathbf{H}^* = 2\nu\text{diag}(1/a_1, 1/a_2)$, $a_j \sim \text{Inverse-Gamma}\left(1/2, 1/A_j^2\right)$ for $j = 1, 2$, $\nu > 0$, and $A_j > 0$ is typically assigned a large value (e.g. 10^5) to indicate non-informativeness. They also showed that the marginal distribution of the correlation coefficient ρ is uniform on $(-1, 1)$ for bivariate cases when $\nu = 2$. Compared to the IW prior, the HIW prior provides increased flexibility in the choice of the scaled matrix while retaining the conjugacy properties. In our model, the conditional posterior distribution of $\mathbf{\Sigma}$ and a_j for $j = 1, 2$ now become

$$p\left(\mathbf{\Sigma}|\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \mu_1, \mu_2, a_1, a_2, \mathbf{x}\right) \propto \text{IW}\left(\nu + I + 1, \mathbf{\Lambda}_\mu + \mathbf{H}^*\right),$$
$$p\left(a_j|\mathbf{\Sigma}, \mathbf{x}\right) \overset{\text{ind}}{\sim} \text{Inverse-Gamma}\left(\tfrac{\nu}{2} + 1, \nu\left(\Sigma^{-1}\right)_{jj} + A_j^{-2}\right),$$

where $\left(\mathbf{\Sigma}^{-1}\right)_{jj}$ denotes the (j, j) entry of $\mathbf{\Sigma}^{-1}$, and we set $\nu = 2$. However, Alvarez, Niemi and Simpson[46] pointed out that, compared to the IW prior, the HIW prior is capable of reducing, but not eliminating, the dependency between variance and correlation.

Separation strategy

Barnard, McCulloch and Meng[45] introduced a prior class known as the separation strategy (SS) that decomposes a covariance matrix $\mathbf{\Sigma}$ into a diagonal matrix $\mathbf{S}$ of standard deviations (SDs) and a correlation matrix $\mathbf{R}$, resulting in $\mathbf{\Sigma} = \mathbf{SRS}$. Specifically, for bivariate data, $\mathbf{S} = \text{diag}(\sigma_1, \sigma_2)$ and $\mathbf{R} = \begin{pmatrix} \rho & 1 \\ 1 & \rho \end{pmatrix}$. The SS prior assigns independent priors for the SDs and correlations, which eliminates the association between variance and correlation, setting it apart from the IW and HIW priors. Posterior computation with the SS prior is usually done via the Hamiltonian Monte Carlo (HMC) algorithm[50], which was later improved by the No-U-Turn sampler[51] in Stan. In the Stan manual[37], the recommended hyperprior settings for the SS prior are $\sigma_j \sim \text{Cauchy}(0, 2.5)$ constrained by $\sigma_j > 0$ for $j = 1, 2$ and $\mathbf{R} \sim \text{LKJCorr}(1)$, where LKJCorr(1) denotes the LKJ prior from Lewandowski et al.[52] with a shape parameter of 1. However, implementing the specific SS prior still requires intensive posterior computation. On the other hand, the IND prior is the simplest among the SS class, which assigns independent priors on $\left(\sigma_1^2, \sigma_2^2, \rho\right)$ for the bivariate case, where $\sigma_j^2 \sim \text{IG}(0.01, 0.01)$ for $j = 1, 2$ and $\rho \sim \text{U}(-1, 1)$ to reflect our lack of information about these terms. The posterior computation involved in the IND prior is much less compared to the SS prior suggested in the Stan manual and can be done via a Gibbs sampler. Thus, throughout our simulation and real data analyses, the IND prior was used for this SS class for computational efficiency.

Scaled inverse Wishart prior

O'Malley and Zaslavsky[44] developed a scaled inverse Wishart (SIW) prior that decomposes a covariance matrix differently such that $\mathbf{\Sigma} = \Delta\mathbf{Q}\Delta$, where for the bivariate case, $\mathbf{Q}$ is an two dimensional unscaled matrix with the (i, j) element $\mathbf{Q}_{ij}$ and $\Delta = \text{diag}(\delta_1, \delta_2)$. The SIW prior is defined as $\mathbf{Q} \sim \text{IW}(\nu, \mathbf{H})$ and $\log\left(\delta_j\right) \overset{\text{ind}}{\sim} \text{N}\left(b_j, \xi_j^2\right)$, which implies that standard deviation $\sigma_j = \delta_j\sqrt{\mathbf{Q}_{jj}}$ for $j = 1, 2$ and $\rho = \frac{\mathbf{Q}_{12}}{\sqrt{\mathbf{Q}_{11}\mathbf{Q}_{22}}}$. Compared to the SS prior, the SIW prior avoids problematic transformation steps, yielding a more efficient sampling process. Following the specifications in Gelman and Hill[53] and Akinc and Vandebroek[47], we set $\mathbf{H} = \mathbf{I}$, $\nu = 3$,$b_j = 0$ and $\xi_j^2 = 1$ for $j = 1, 2$.

Simulation

We conducted two simulation studies focusing on meta-analysis of rare binary events: the first is to compare the performance of our Bayesian model under various covariance prior choices on the estimation of key model parameters in terms of bias, mean squared error (MSE) and coverage; the second is to assess the performance of our proposed IPQ method in comparison to existing GOF tests in terms of Type I error rates and statistical power. Our code is publicly available at https://github.com/chriszhangm/MetaGOF. In our numerical experiments, we used the default continuity correction factor of 0.5 for all frequentist methods unless otherwise stated. On the other hand, we did not adopt any continuity correction or eliminate studies containing zero events for Bayesian methods since they handle such studies automatically via incorporation of prior information into data analysis.

Comparison of different covariance prior choices

We simulated data using the generalized REM in (2) to evaluate the performance of our Bayesian model with four covariance prior choices (i.e., IW, HIW, SIW and IND) in estimating (a) the overall treatment effect θ_0, (b) the heterogeneity τ^2 and (c) the correlation coefficient ρ , based on three metrics (bias, MSE and coverage). Specifically, for a parameter of interest (say κ, κ can be θ_0, τ^2 or ρ), we define $\text{Bias}(\kappa) = \frac{\sum_{j=1}^{J}(\hat{\kappa}_j - \kappa)}{J}$ and $\text{MSE}(\kappa) = \frac{\sum_{j=1}^{J}(\hat{\kappa}_j - \kappa)^2}{J}$, where $\hat{\kappa}_j$ is the corresponding estimate of κ in the j^{th} replicated dataset. We estimated θ_0 and ρ using the posterior mean, and τ^2 using the posterior median due to its heavily skewed distribution. The coverage probability was computed using 95% equal-tail credible intervals.

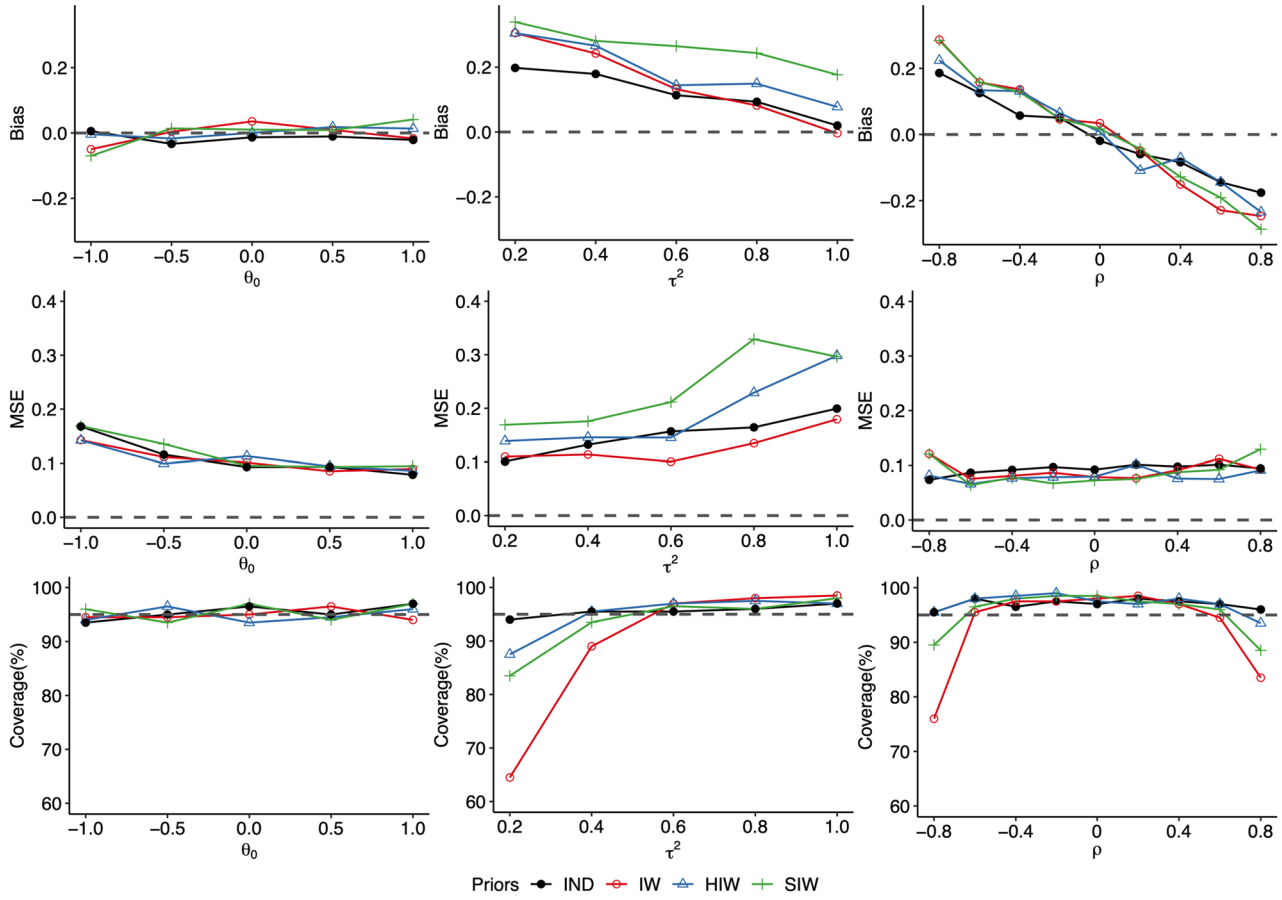

Figure 2. Comparison of bias, MSE and coverage results for Bayesian estimates of θ_0, τ^2 and ρ using the Inverse-Wishart prior (IW), the Hierarchical Inverse-Wishart prior (HIW), the Scaled Inverse-Wishart prior (SIW) and the independent prior (IND) for meta-analysis of rare binary events with $I = 20$ studies. For each setting, results were generated using 200 replicates and the nominal coverage level was set as 95%.

To generate data from (2), we set $\mu_1 = -5$, $\mu_2 = -5 + \theta_0$, and $\sigma_1^2 = \sigma_2^2 = 0.5$. For (a), we varied $\theta_0 \in \{-1, -0.5, 0, 0.5, 1\}$ and fixed $\rho = 0$ so that $\tau^2 = 1$ (recall that $\tau^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$) while for (b), we varied $\tau^2 \in \{0.2, 0.4, ..., 1\}$ so that $\rho = 1 - \tau^2$, and fixed $\theta_0 = 0$; for (c), we varied $\rho = \{-0.8, -0.6, ..., 0.6, 0.8\}$ so that $\tau^2 = 1 - \rho$, and fixed $\theta_0 = 0$. Then, for each setting, we simulated probabilities $p_{i1} = \frac{\exp(\phi_{i1})}{1+\exp(\phi_{i1})}$ and $p_{i2} = \frac{\exp(\phi_{i2})}{1+\exp(\phi_{i2})}$ for study $i = 1, ..., I$. We considered three meta-analysis sizes $I = 20, 50, 80$, and allowed different sample-size allocation ratios across studies by setting $n_{i2} = r_i n_{i1}$, where $\log_2 r_i \sim \mathrm{N}(0, 0.5)$. The number of subjects in the control group n_{i1}, was randomly drawn from 50 to 1000 for each study, and the number of events x_{i1} or x_{i2} was generated by $\mathrm{Bin}(n_{i1}, p_{i1})$ or $\mathrm{Bin}(n_{i2}, p_{i2})$.

In Figure 2, we report results based on 200 replicates for each setting with $I = 20$, in which the three rows give bias, MSE and coverage results, and the three columns correspond to θ_0, τ^2 and ρ, respectively. We observe that the performance of IPQ in estimating θ_0 seems to be insensitive to the choices of different priors. For τ^2 and ρ, the IND prior generally outperforms other choices as it produces smaller bias and MSE as well as coverage closer to the nominal level 95% in most scenarios. Among the other three priors, although IW tends to do better in point estimation, it gives the worst coverage for both τ^2 and ρ especially when τ^2 is small or $|\rho|$ is close to 1. This is not surprising, as mentioned before, the IW prior induces dependency between variance and correlation, which can bias the inference. For $I = 50$ or 80 (results omitted here for brevity), while the discrepancies of the bias and coverage results using different priors become less salient, it is worth noting that IW still yields unsatisfactory coverage.

For GOF testing using the proposed IPQ method in this paper, the IND prior was used due to its demonstrated better performance and its simplicity.

Performance evaluation of GOF testing

Our interest lies in conducting the GOF test to detect departures from a common assumption in meta-analysis that the true effect sizes θ_is of component studies are normally distributed. Using the generalized REM, for null cases, we generated ϕ_{i1} from normal distributions; for non-null cases, we generated ϕ_{i1} from one of four pre-specified non-normal distributions: (i) an exponential distribution with a rate of 1; (ii) a gamma distribution with a shape parameter of 4 and a scale parameter of 0.5, a unimodal right-skewed distribution; (iii) a mixture of

two equal-weighted normal distributions: N(1, 0.5) and N(4, 0.5); (iv) a t distribution with number of degrees of freedom of 4. Then, with $z \sim \mathrm{N}(0,1)$, ϕ_{i2} given ϕ_{i1} was generated by $\phi_{i2} = \rho\frac{\sigma_2}{\sigma_1}\phi_{i1} + \sigma_2\sqrt{1-\rho^2}z + \mu_2 - \rho\frac{\sigma_2}{\sigma_1}\mu_1$ so that the correlation coefficient between (ϕ_{i1}, ϕ_{i2}) is ρ and the mean and variance of ϕ_{ij} are μ_j and σ_j^2 for $j = 1, 2$, respectively. In other words, the conditional distribution of ϕ_{i2} given ϕ_{i1} is set to be normal in our simulation. Note that the four distributions of ϕ_{i1} cover different types of violation of the normality assumption, of which the first two cover skewness, the next covers multimodality, and the last covers heavy-tailedness. Generating the data in the above way would pass these types of violation to the distribution of θ_i's.

Without loss of generality, we set the means of ϕ_{i1} and ϕ_{i2} to be equal such that the overall effect $\theta_0 = 0$, $\mu = \mu_1 = \mu_2 \in \{-5, -3, -2\}$, corresponding to $\{0.67\%, 4.74\%, 11.92\%\}$ in probability scale. We set $\sigma_1^2 = 0.5, \sigma_2^2 = 0.8$ for the null cases , and set $\sigma_2^2 = 0.8$ while σ_1^2 was determined by its distribution type specified above for the non-null cases(e.g., $\sigma_1^2 = 1$ for ϕ_{i1} from the exponential distribution). We further set the number of component studies $I \in \{20, 50, 80\}$, and $\rho \in \{-0.5, 0, 0.5\}$.

We compared our proposed method IPQ to six other approaches, including three frequentist-based approaches: the Naïve method that conducts the Shapiro-Wilk test on estimated effect sizes (log odds ratios) directly, the parametric bootstrap method (PB[8]) and the standardization method (STD[9]), and three Bayesian methods: the pivotal quantities method (PQ[10, 11]) using two cutoffs of 0.25 and 0.1, the posterior predictive check (PPC[15]) using the discrepancy function recommended in Sinharay and Stern[54], defined as $T(\boldsymbol{\theta}) = |\max(\boldsymbol{\theta}) - \mathrm{median}(\boldsymbol{\theta})| - |\min(\boldsymbol{\theta}) - \mathrm{median}(\boldsymbol{\theta})|$ with $\boldsymbol{\theta} = (\theta_i)_{i=1}^I$, and the sampled posterior check (SPC[21, 22]) using the same discrepancy function. We set the significance level $\alpha = 0.05$. For all frequentist approaches (PB, STD and Naive) and the proposed IPQ, we reject the normality assumption if the p value is less than 0.05. For PPC or SPC, we reject the null when the posterior predictive p value (PPP) is below 0.025 or above 0.975[22]; for PQ, if the minimum p value upper bound $p_{\min}$ is less than the chosen cutoff (0.25 or 0.1), we reject the null. As mentioned earlier, for either PPC or PQ, the reference distribution of PPP or $p_{\min}$ is not uniform(0,1) even in an asymptotic sense, and so we do not expect that they maintain the Type I error rate. However, the cutoff value 0.1 for $p_{\min}$ in the PQ method was chosen via preliminary simulation because it can offer error rates much closer to 0.05 in most of the simulation scenarios, compared to the rule-of-thumb value

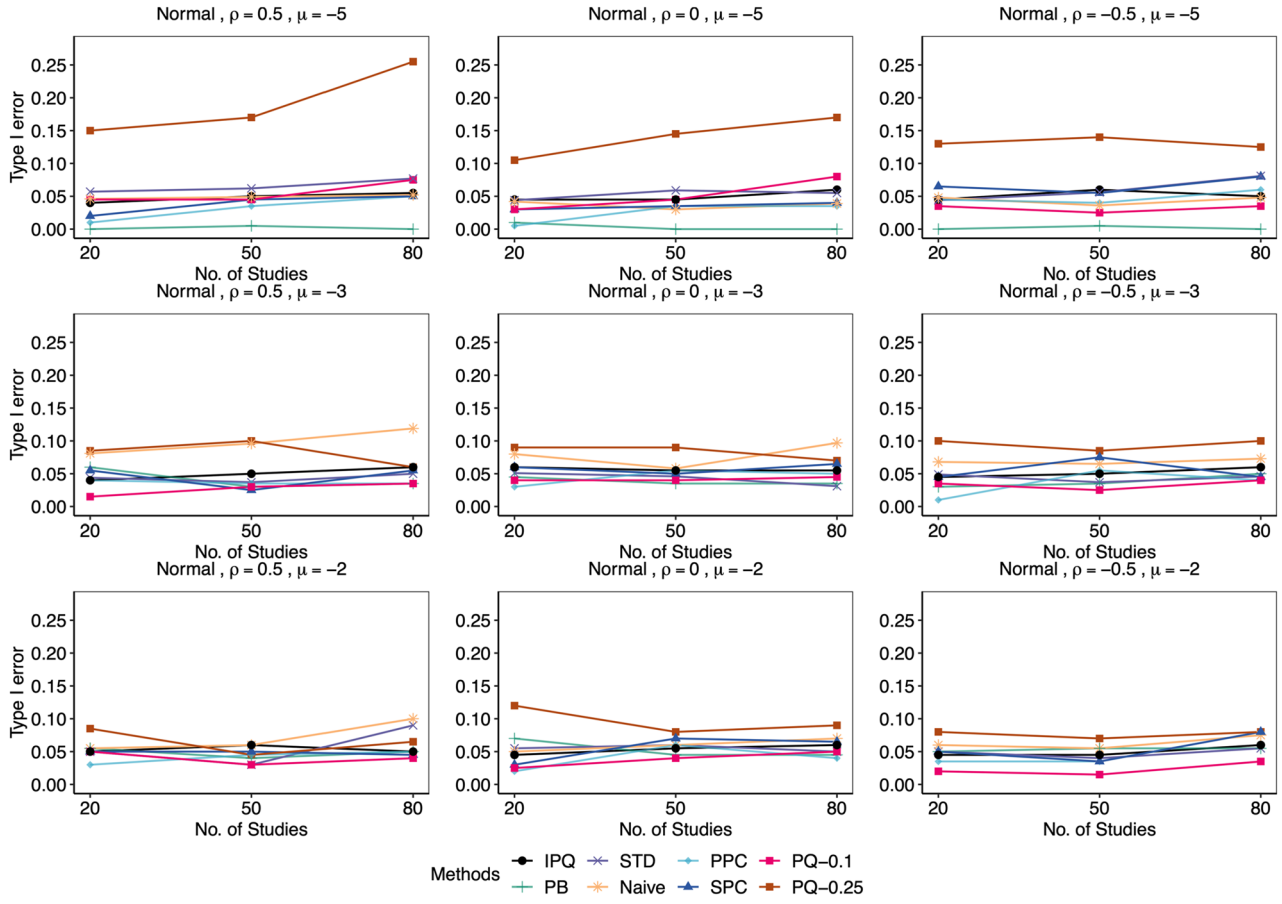

Figure 3. Comparison of empirical Type I error rates by proposed IPQ, parametric bootstrap method (PB), standardization method (STD), Naïve method (Naive), Posterior Predictive Check (PPC), Sampled Posterior Check (SPC), and Pivotal Quantities method with cutoffs of 0.1 and 0.25 (PQ-0.1, PQ-0.25). Data were generated from the null cases with different (I, μ, ρ) combinations, each 200 replicates. Tests were conducted at the significant level $\alpha = 0.05$.

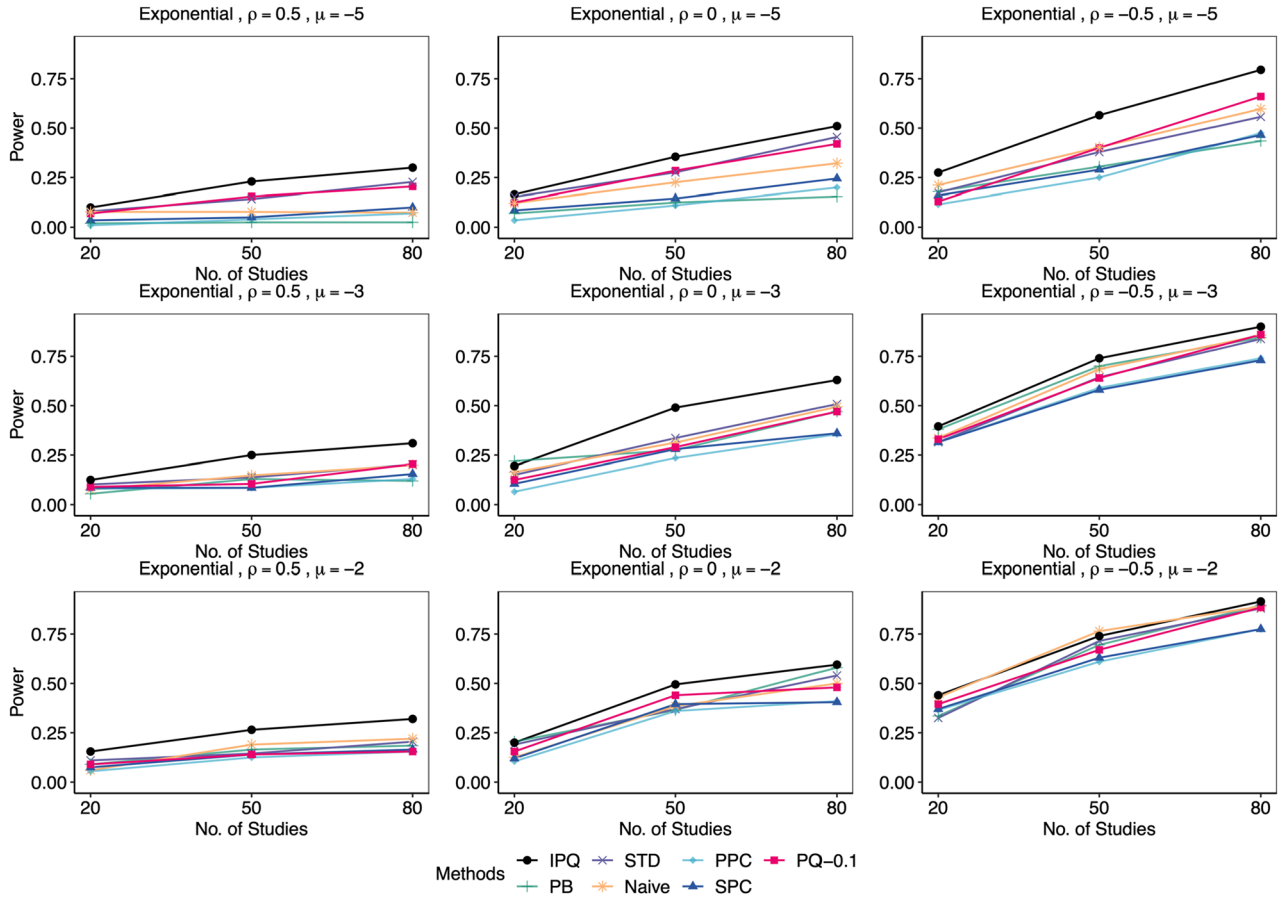

Figure 4. Comparison of empirical power by proposed IPQ, parametric bootstrap method (PB), standardization method (STD), Naïve method (Naive), Posterior Predictive Check (PPC), Sampled Posterior Check (SPC), and Pivotal Quantities method with a cutoff of 0.1(PQ-0.1). Data were generated from the non-null cases, where $\phi_{i1} \sim \text{Exp}(1)$, with different (I, μ, ρ) combinations, each 200 replicates. Tests were conducted at the significant level $\alpha = 0.05$.

of 0.25. We simulated 200 replicates for each setting, and reported Type I error rates for data from null cases and statistical power otherwise.

Figure 3 reports the Type I error rates for all methods in various settings. The IPQ, STD, and SPC methods demonstrate superior performance, as they maintain an error rate close to the nominal value of 0.05 regardless of $\{I, \mu, \rho\}$. Conversely, PQ-0.25 (the PQ method with the recommended cutoff of 0.25[11]) frequently produces severely inflated Type I error rates, particularly as the event of interest becomes rarer, while PQ-0.1 performs much better in general. Therefore, we exclude PQ-0.25 from our power results below. Among the remaining three methods, PPC and PB are often conservative for rarer events (i.e., $\mu = -5$), exhibiting Type I error rates below 0.05, while Naive tends to have inflated rates for less rare events.

Figures 4, 5, 6, 7 display power results with different underlying distributions of ϕ_{i1}. We observe that all approaches tend to report higher power as I increases or ρ decreases. Also, the differences in power among the methods become smaller as μ goes up. This is perhaps because some methods in the bottom group such as PB improve significantly while the proposed IPQ, as the best overall method, appears to be much less sensitive to the change of μ. Figures 4 and 5 present power results for skewed distributions (i.e., exponential and gamma distributions). IPQ is the best in nearly all scenarios, followed by PQ-0.1 and STD. Naive often stands somewhere in the middle among all. PB tends to perform poorly, except for larger μ and smaller ρ. SPC reports slightly better results than PPC, while both methods provide the worst overall results, particularly with large ρ. Figure 6 presents outcomes for a multimodal distribution (i.e., normal mixture). IPQ is a clear winner and provides the highest power in all settings. Among the others, STD and PQ-0.1 usually perform better, followed by Naïve and PB. SPC and PPC consistently give the worst results that show almost no power. Figure 7 displays power results for t_4, a symmetric and heavy-tailed distribution, where again, IPQ outperforms other methods while PPC and SPC tend to perform the worst.

To summarize, the proposed IPQ maintains the Type I error rate at the target level well and offers the highest statistical power for various departures from the normality assumption compared to alternative approaches.

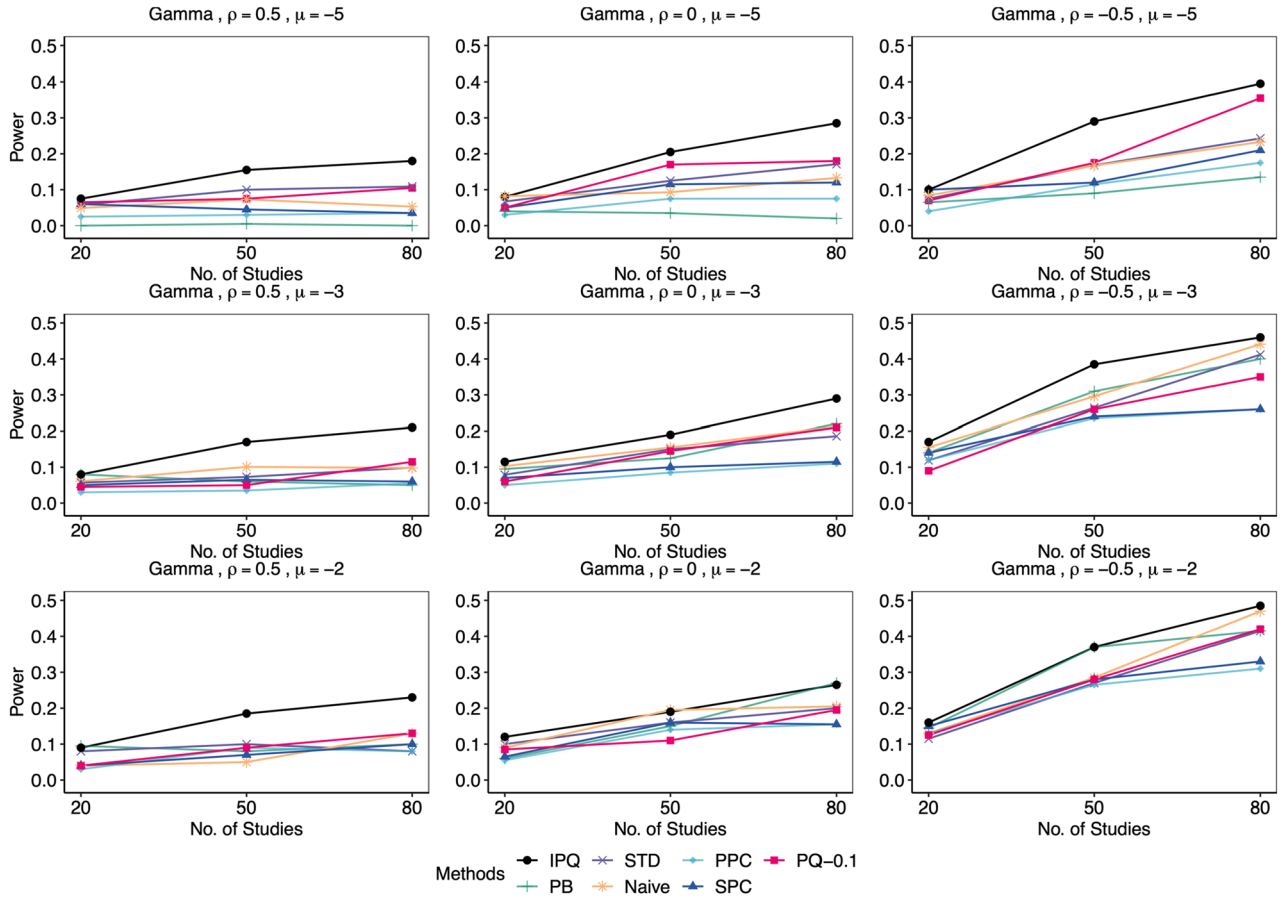

Figure 5. Comparison of empirical power by proposed IPQ, parametric bootstrap method (PB), standardization method (STD), Naïve method (Naive), Posterior Predictive Check (PPC), Sampled Posterior Check (SPC), and Pivotal Quantities method with cutoff 0.1(PQ-0.1). Here, data were generated from the non-null cases, where $\phi_{i1} \sim \text{Gamma}(4, 0.5)$, with different (I, μ, ρ) combinations, each 200 replicates. Tests were conducted at the significant level $\alpha = 0.05$.

Data examples

We applied our IPQ method, along with six other methods (PB, STD, Naïve, PPC, SPC and PQ-0.1) to three real data sets of meta-analysis, for testing the normality assumption about the distribution of true effect sizes across component studies. The first involves hand-eye dominance data, the second involves diabetes data, and the third involves lung cancer data.

Bourassa[55] conducted a meta-analysis of 54 studies to investigate the hand-eye dominance association (see Table A1 for detailed data in Supplementary Material). The study found that the hand-eye concordance was larger than one, indicating left-handed people tended to have left-eyed dominance, and the same was true for right-handed people. The meta-analysis included 54,087 subjects, summarized in 2×2 tables of four categories: left-handed/left-eyed, left-handed/right-eyed, right-handed/left-eyed and right-handed/right-eyed. We considered the event of interest to be " left-handed," with the control and case groups being " left-eyed" and " right-eyed," respectively. The overall incident rates for the control and case groups are about 6% and 18.5%, which are −2.75 and −1.48 on a logit scale equivalently. The left panel of Figure 8 displays the histogram, density and quantile-quantile plots of the observed log odds ratio, revealing a left-skewed distribution.

Bellamy et al.[56] conducted a meta-analysis of 20 studies to investigate the association between Type 2 diabetes mellitus and gestational diabetes (see Table A2 for detailed data in Supplementary Material). The analysis revealed that women with gestational diabetes had an increased risk of developing type 2 diabetes. The study included 675,455 subjects, of which 31,867 had Type 2 diabetes. Among the control groups (no gestational diabetes), 6,862 subjects had Type 2 diabetes, indicating an overall incident rate of ~1.1% (or −4.53 on a logit scale). For the case groups (with gestational diabetes), 3997 of them had Type 2 diabetes, resulting in an incident rate of ~12.5% (or −1.94 on a logit scale). The middle panel of Figure 8 shows the histogram, density, and quantile-quantile plots of the observed log odds ratio, suggesting a unimodal, symmetric, and bell-shaped curve.

Feng et al.[57] conducted a meta-analysis of 44 studies to evaluate the association between GSTP1 gene polymorphism and the risk of lung cancer (see Table A3 for detailed data in Supplementary Material). The event of interest is considered the GG genotype of GSTP1. The study included 26,516 subjects, of which 2763 had the GG genotype. Among the control (no lung cancer) and case (lung cancer) groups, 1406 and 1357 subjects had the GG genotype, implying overall incident rates of 10.0 (or −2.19 on a logit scale) and 10.8 (or −2.11 on a logit

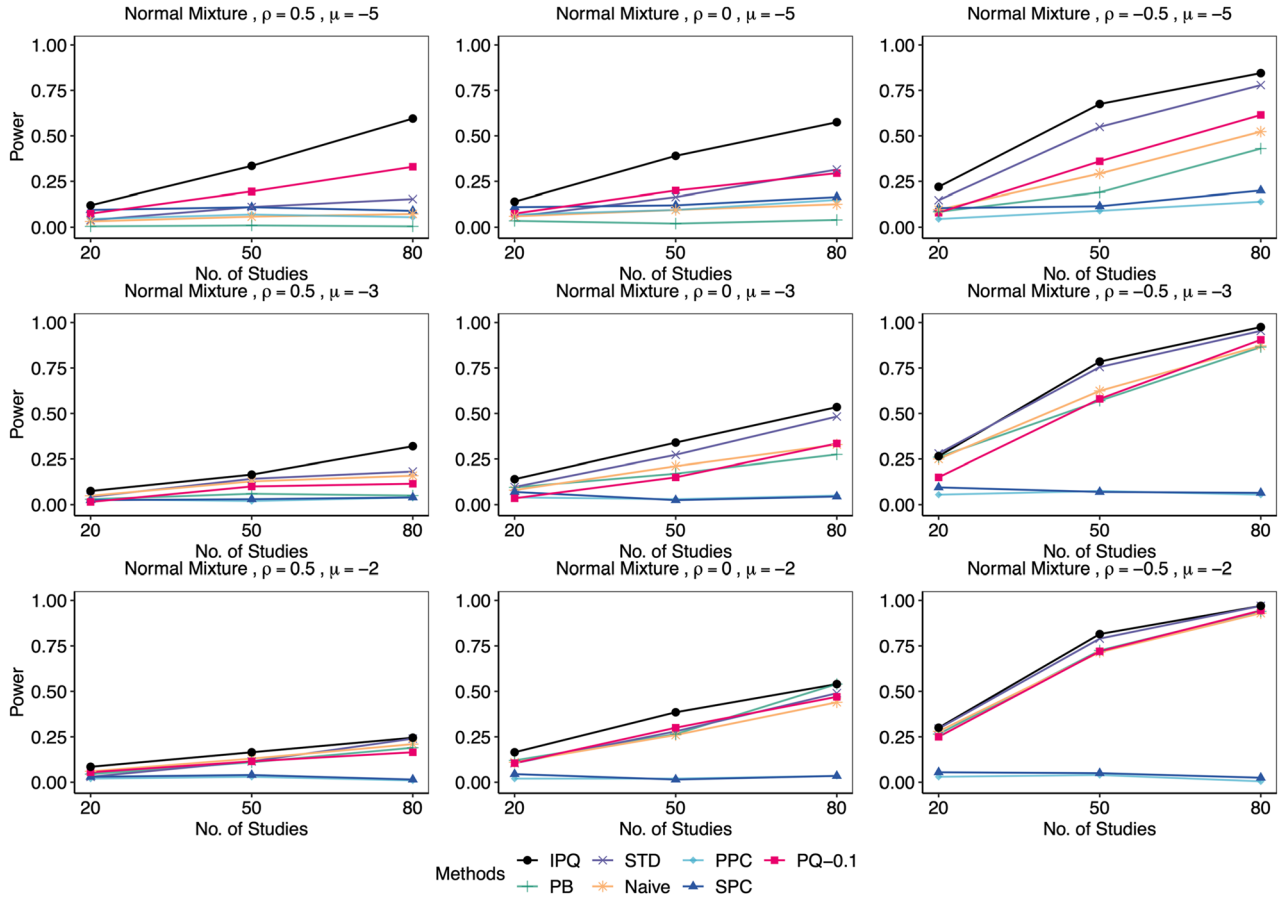

Figure 6. Comparison of empirical power by proposed IPQ, parametric bootstrap method (PB), standardization method (STD), Naïve method (Naive), Posterior Predictive Check (PPC), Sampled Posterior Check (SPC), and Pivotal Quantities method with a cutoff of 0.1(PQ-0.1). Here, data were generated from the non-null cases, where $\phi_{i1} \sim 0.5\mathrm{N}(1, 0.5) + 0.5\mathrm{N}(4, 0.5)$, with different (I, μ, ρ) combinations, each 200 replicates. Tests were conducted at the significant level $\alpha = 0.05$.

scale), respectively. The right panel of Figure 8 reveals the histogram, density, and quantile-quantile plots of the observed log odds ratio, showing a roughly symmetric curve but with heavy tails on both sides.

Table 1 shows that for the hand-eye dominance data, at the significance level $\alpha = 0.05$, all methods except for STD reject the null hypothesis, indicating a departure from the assumed normality. Note that among the existing methods, STD was quite competitive. Nevertheless, it failed in this specific example. On the other hand, PPC and SPC tend to be conservative in rejecting the null but worked here. For the diabetes data, Table 1 shows that all methods have the same conclusion: there is no evidence against the normality.

For the GSTP1 gene polymorphism and lung cancer data, Table 1 shows that Naïve, PQ-0.1 and IPQ reject the null hypothesis while other methods do not provide evidence against normality. As shown in Figure 7, given the symmetric and heavy-tailed distribution, IPQ offered the highest power across all the cases. When $\mu = -2$, similar to the overall incidence rates in this dataset, Naïve and PQ-0.1 performed relatively well. However, STD, PPC, and SPC performed poorly under the cases of $\mu = -2$. Therefore, we recommend avoiding the normality assumption in this example.

In summary, IPQ performs consistently well across all three real data examples, while PB, STD, PPC, and SPC sometimes fail. Although Naïve and PQ-0.1 also demonstrate good performance here, they are less satisfactory in our simulation studies. As such, IPQ is the recommended method of choice in this context.

Discussion

Meta-analysis commonly assumes that actual effect sizes from component studies follow a normal distribution for mathematical convenience, despite a lack of formal justification for this assumption. In practice, however, this assumption can be frequently violated, potentially leading to inaccurate conclusions. To address this issue, we propose a novel goodness-of-fit (GOF) test called Improved Pivotal Quantities (IPQ) for testing this assumption in the context of meta-analysis of rare binary outcomes, where the effect size is measured by log odds ratio.

The proposed IPQ method builds upon the strengths of the original PQ approach[10], which is conceptually simple and efficient in detecting model misfit at any level of a hierarchical model without additional computational costs. However, the original PQ method employs the probability bound as a criterion to determine model misfit, which can result in inflated Type I error rates when used with the rule-of-thumb cutoff of 0.25. This highlights

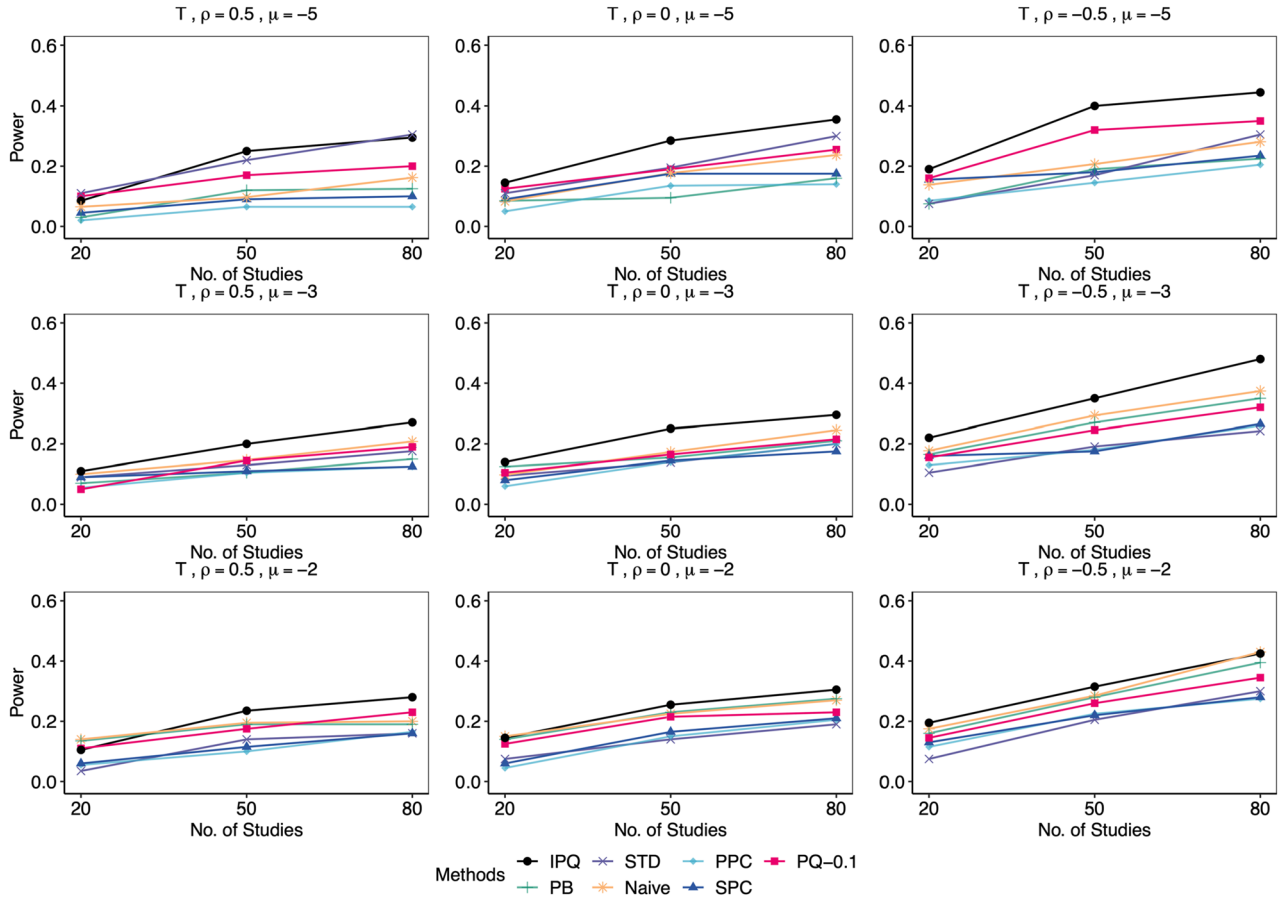

Figure 7. Comparison of empirical power by proposed IPQ, parametric bootstrap method (PB), standardization method (STD), Naïve method (Naive), Posterior Predictive Check (PPC), Sampled Posterior Check (SPC), and Pivotal Quantities method with a cutoff of 0.1(PQ-0.1). Here, data were generated from the non-null cases, where $\phi_{i1} \sim t_4$, with different (I, μ, ρ) combinations, each 200 replicates. Tests were conducted at the significant level $\alpha = 0.05$.

the need for selecting new cutoff values that are tailored to different applications. To address this limitation, our IPQ method improves the decision-making process of PQ by adopting the Cauchy combination idea[12] to account for dependent p value. In addition, given sparse data such as tables with zero events, IPQ naturally incorporates all data, without requiring artificial corrections due to its Bayesian model formulation.

In fact, IPQ is a hybrid approach. It adopts the frequentist framework for hypothesis testing, since it uses the Cauchy combination test to obtain a p value, from which the final conclusion is drawn. On the other hand, it constructs the test statistics by incorporating a Bayesian idea through Markov Chain Monte Carlo methods. We further note that, because of the use of pivotal quantities, the sampling distribution of the proposed test statistics, evaluated at posterior samples, is known and invariant (i.e., N(0,1)) under the null hypothesis. The set of posterior draws used to construct the test statistics is from the same data (i.e., the true observed data rather than any "fake" data).

Simulation results indicate that IPQ maintains well-controlled Type I error rates while achieving higher statistical power than other approaches in most scenarios. To demonstrate the effectiveness of our method, we provide examples of three real datasets. Specifically, our results suggest that the normality assumption should be avoided for the hand-eye dominance dataset[55] and the GSTP1 gene polymorphism and lung cancer dataset[57], while it is likely to hold for the diabetes dataset[56]. In situations where the normality assumption does not hold, it becomes imperative to explore alternative distributions, such as those characterized by heavy tails (e.g., t

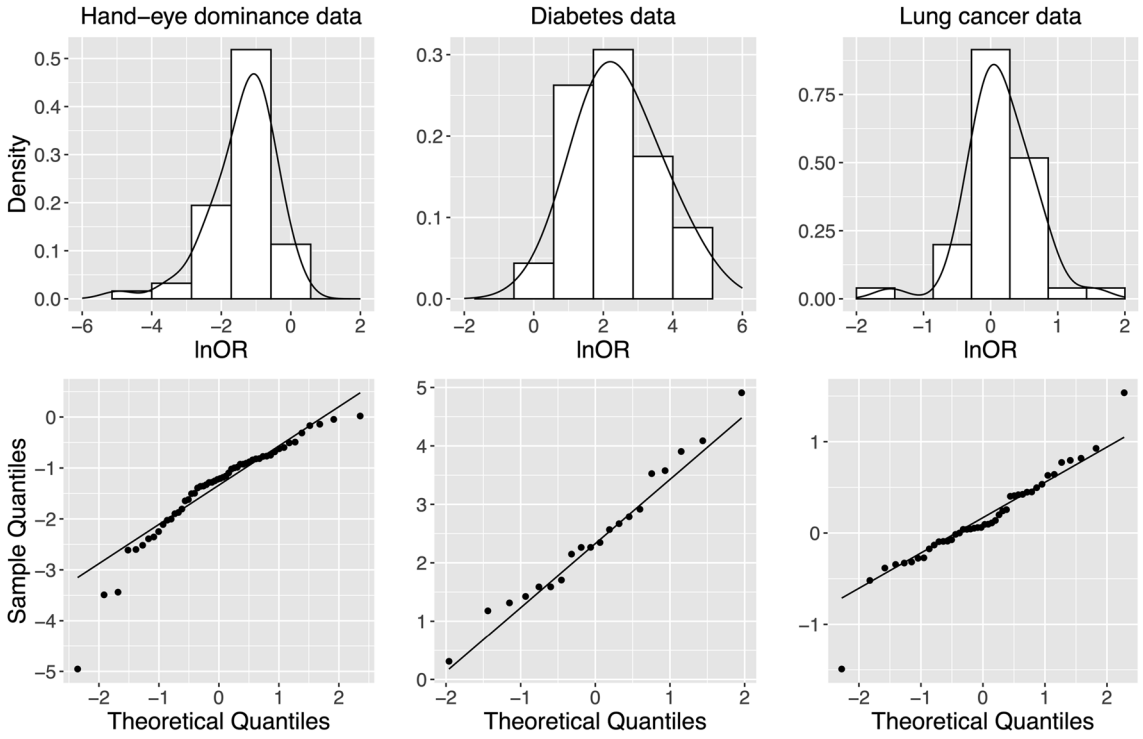

Figure 8. The histogram and density plots (top) and Quantile-Quantile plot (bottom) of the observed effect sizes measured by log odds ratio (lnOR). The left panel is for hand-eye dominance data, the middle panel is for diabetes data and the right panel is for lung cancer data.

Methods	Hand-eye dominance	Diabetes	Lung cancer
PB	0.021	0.932	0.753
STD	0.069	0.835	0.460
Naïve	<0.001	0.921	0.034
PPC	0.999	0.277	0.666
SPC	0.998	0.245	0.714
PQ-0.1	0.031	0.707	0.022
IPQ	0.007	0.985	0.021

Table 1. *P* values of the GOF tests for three meta-analyses involving (i) hand-eye dominance data, (ii) type 2 diabetes mellitus and gestational diabetes data, and (iii) GSTP1 gene polymorphism and lung cancer data. Note that for PPC or SPC, the posterior predictive *p* value is reported and for PQ-0.1, the minimum *p* value upper bound $p_{\min}$ is reported.

distributions) or skewness (e.g., gamma distributions), in order to more accurately capture the characteristics of observed data. Alternatively, one can employ nonparametric methods for estimating treatment effects[58] and for estimating heterogeneity[59–61]. Furthermore, in scenarios where a meta-analysis involves a small number of studies, a situation commonly encountered in practice, alternative frameworks such as Bayesian model averaging may yield more reliable outcomes.

Although our focus is primarily on rare binary events, the IPQ method is directly applicable to meta-analysis of any binary data. However, we believe that the gain in performance for common binary events may not be as significant as that for rare binary events. As demonstrated in our simulation studies, the differences in power between our method and other approaches diminish when increasing the background incidence rate. Moreover, IPQ can be extended beyond testing normality to other scenarios where an appropriate test statistic can be designed to measure the discrepancy. In conclusion, our IPQ method is useful for detecting model misfits and selecting appropriate statistical models for different applications, particularly in scenarios where sparse data are present or when the normality assumption is in question.

Data availability

The data that support the findings of this study are included in Supplementary Material.

Received: 10 July 2023; Accepted: 10 October 2023
Published online: 18 October 2023

References

1. Borenstein, M., Hedges, L. V., Higgins, J. P. & Rothstein, H. R. *Introduction to meta-analysis* (John Wiley & Sons, 2021).
2. Bhaumik, D. K. *et al.* Meta-analysis of rare binary adverse event data. *J. Am. Stat. Assoc.* **107**, 555–567 (2012).
3. Smith, T. C., Spiegelhalter, D. J. & Thomas, A. Bayesian approaches to random-effects meta-analysis: A comparative study. *Stat. Med.* **14**, 2685–2699 (1995).
4. Li, L. & Wang, X. Meta-analysis of rare binary events in treatment groups with unequal variability. *Stat. Methods Med. Res* **28**, 263–274 (2017).
5. Houwelingen, H. C. V., Zwinderman, K. H. & Stijnen, T. A bivariate approach to meta-analysis. *Stat. Med.* **12**, 2273–2284 (1993).
6. Lee, K. J. & Thompson, S. G. Flexible parametric models for random-effects distributions. *Stat. Med.* **27**, 418–434 (2008).
7. Wang, C.-C. & Lee, W.-C. A simple method to estimate prediction intervals and predictive distributions: Summarizing meta-analyses beyond means and confidence intervals. *Res. Synth. Methods* **10**, 255–266 (2019).
8. Chen, Z., Zhang, G. & Li, J. Goodness-of-fit test for meta-analysis. *Sci. Rep.* **5**, 456213 (2015).
9. Wang, C.-C. & Lee, W.-C. Evaluation of the normality assumption in meta-analyses. *Am. J. Epidemiol.* **189**, 235–242 (2019).
10. Johnson, V. E. Bayesian model assessment using pivotal quantities. *Bayesian Anal.* **2**, 719–733 (2007).
11. Yuan, Y. & Johnson, V. E. Goodness-of-fit diagnostics for bayesian hierarchical models. *Biometrics* **68**, 156–164 (2011).
12. Liu, Y. & Xie, J. Cauchy combination test: A powerful test with analytic *p* value calculation under arbitrary dependency structures. *J. Am. Stat. Assoc.* **115**, 393–402 (2019).
13. Box, G. E. P. Sampling and bayes'inference in scientific modelling and robustness. *J. R. Stat. Ser. A General* **143**, 383–430 (1980).
14. Stern, H. S. *Handbook of Statistics* (Elsevier, 2005).
15. Gelman, A., Li Meng, X. & Stern, H. Posterior predictive assessment of model fitness via realized discrepancies. *Stat. Sin.* **56**, 733–807 (1996).
16. Bayarri, M. J. & Castellanos, M. E. Bayesian checking of the second levels of hierarchical models. *Stat. Sci.* **22**, 322–343 (2007).
17. Dey, D. K., Gelfand, A. E., Swartz, T. B. & Vlachos, P. K. A simulation-intensive approach for checking hierarchical models. *Test* **7**, 325–346 (1998).
18. Hjort, N. L., Dahl, F. A. & Steinbakk, G. H. Post-processing posterior predictive p values. *J. Am. Stat. Assoc.* **101**, 1157–1174 (2006).
19. Bayarri, M. J. & Berger, J. O. P values for composite null models. *J. Am. Stat. Assoc.* **95**, 1127–1142 (2000).
20. Johnson, V. E. Comment: Bayesian checking of the second levels of hierarchical models. *Stat. Sci.* **22**, 353–358 (2007).
21. Gosselin, F. A new calibrated bayesian internal goodness-of-fit method: Sampled posterior *p* values as simple and general *p* values that allow double use of the data. *PloS one* **6**, e14770 (2011).
22. Zhang, J. L. Comparative investigation of three bayesian p values. *Comput. Stat. Data Anal.* **79**, 277–291 (2014).
23. Gascuel, O. & Caraux, G. Bounds on expectations of order statistics via extremal dependences. *Stat. Probab. Lett.* **15**, 143–148 (1992).
24. Rychlik, T. Stochastically extremal distributions of order statistics for dependent samples. *Stat. Probab. Lett.* **13**, 337–341 (1992).
25. Li, L., Wu, T. & Feng, C. Model diagnostics for censored regression via randomized survival probabilities. *Stat. Med.* **40**, 1482–1497 (2020).
26. Zhang, C., Wang, X., Chen, M. & Wang, T. A comparison of hypothesis tests for homogeneity in meta-analysis with focus on rare binary events. *Res. Synth. Methods* **12**, 408–428 (2021).
27. Zhang, M., Barth, J., Lim, J. & Wang, X. Bayesian estimation and testing in random-effects meta-analysis of rare binary events allowing for flexible group variability. *Stat. Med.* **42**, 1699–1721 (2023).
28. Fisher, R. A. Statistical methods for research workers. In *Springer Series in Statistics*, 66–70 (Springer, 1992).
29. Tippett, L. *The methods of statistics* (1931).
30. Berk, R. H. & Jones, D. H. Goodness-of-fit test statistics that dominate the kolmogorov statistics. *Z. Wahrscheinlichkeitstheorie Verwandte Geb.* **47**, 47–59 (1979).
31. Donoho, D. & Jin, J. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Stat.* **32**, 962–994 (2004).
32. Shapiro, S. S. & Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika* **52**, 591–611 (1965).
33. Cramér, H. On the composition of elementary errors: First paper: Mathematical deductions. *Scand. Actuar. J.* **1928**, 13–74 (1928).
34. Mises, R. V. *Wahrscheinlichkeit Statistik und Wahrheit* (Springer-Verlag, London, 2013).
35. D'Agostino, R. *Goodness-of-Fit-Techniques* (Routledge, 2017).
36. Anderson, T. W. & Darling, D. A. A test of goodness of fit. *J. Am. Stat. Assoc.* **49**, 765–769 (1954).
37. Stan Development Team, . Stan modeling language users guide and reference manual, 2.29. *"* (2022).
38. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing (2019).
39. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **25**, 457–472 (1992).
40. Bai, O., Chen, M. & Wang, X. Bayesian estimation and testing in random effects meta-analysis of rare binary adverse events. *Stat. Biopharm. Res.* **8**, 49–59 (2016).
41. Rúa, S. M. H., Mazumdar, M. & Strawderman, R. L. The choice of prior distribution for a covariance matrix in multivariate meta-analysis: a simulation study. *Stat. Med.* **34**, 4083–4104 (2015).
42. Berger, J. O. *Stat. Decis. Theory Bayesian Anal.* (Springer, New York, 1985).
43. Huang, A. & Wand, M. P. Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Anal.* **8**, 439–452 (2013).
44. O'Malley, A. J. & Zaslavsky, A. M. Domain-level covariance analysis for multilevel survey data with structured nonresponse. *J. Am. Stat. Assoc.* **103**, 1405–1418 (2008).
45. Barnard, J., McCulloch, R. & Meng, X. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Stat. Sin.* **10**, 1281–1311 (2000).
46. Alvarez, I., Niemi, J. & Simpson, M. Bayesian inference for a covariance matrix. *Ann. Conf. Appl. Stat. Agric.* **26**(2014), 71–82 (2014).
47. Akinc, D. & Vandebroek, M. Bayesian estimation of mixed logit models: Selecting an appropriate prior for the covariance matrix. *J. Choice Modell.* **29**, 133–151 (2018).
48. Wand, M. P., Ormerod, J. T., Padoan, S. A. & Frühwirth, R. Mean field variational bayes for elaborate distributions. *Bayesian Anal.* **6**, 847–900 (2011).
49. Armagan, A., Dunson, D. B. & Clyde, M. Generalized beta mixtures of gaussians. *Adv. Neural Inf. Process. Syst.* **24**, 523–531 (2011).
50. Duane, S., Kennedy, A., Pendleton, B. J. & Roweth, D. Hybrid monte carlo. *Phys. Lett. B* **195**, 216–222 (1987).
51. Hoffman, M. D. *et al.* The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.* **15**, 1593–1623 (2014).
52. Lewandowski, D., Kurowicka, D. & Joe, H. Generating random correlation matrices based on vines and extended onion method. *J. Multivar. Anal.* **100**, 1989–2001 (2009).
53. Gelman, A. & Hill, J. *Data analysis using regression and multilevel/hierarchical models* (Cambridge University Press, 2006).
54. Sinharay, S. & Stern, H. S. Posterior predictive model checking in hierarchical models. *J. Stat. Plan. Inference* **111**, 209–221 (2003).

55. Bourassa, D. Handedness and eye-dominance: A meta-analysis of their relationship. *Laterality* **1**, 5–34 (1996).
56. Bellamy, L., Casas, J.-P., Hingorani, A. D. & Williams, D. Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. *Lancet* **373**, 1773–1779 (2009).
57. Feng, X. *et al.* Association of glutathione s-transferase p1 gene polymorphism with the susceptibility of lung cancer. *Mol. Biol. Rep.* **39**, 10313–10323 (2012).
58. DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Controll. Clin. Trials* **7**, 177–188 (1986).
59. Malzahn, U., Böhning, D. & Holling, H. Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika* **87**, 619–632 (2000).
60. Kontopantelis, E., Springate, D. A. & Reeves, D. A re-analysis of the cochrane library data: the dangers of unobserved heterogeneity in meta-analyses. *PloS one* **8**, e69930 (2013).
61. Wang, G., Cheng, Y., Chen, M. & Wang, X. Jackknife empirical likelihood confidence intervals for assessing heterogeneity in meta-analysis of rare binary event data. *Contemp. Clin. Trials* **107**, 106440 (2021).

Acknowledgements

This study was supported by the National Institutes of Health (Grant No.: R15GM131390 to X. Wang).

Author contributions

X.W. conceived, guided, and designed the study. M.Z. and X.W. developed the IPQ method and wrote the first draft of the manuscript. M.Z. implemented the algorithm, wrote computer code, conducted simulation studies and data analysis. J.L. and O.X. revised the draft and provided helpful discussions and feedback. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-44638-x.

Correspondence and requests for materials should be addressed to X.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023