Received: 22 November 2019 | Revised: 10 August 2020 | Accepted: 18 August 2020

DOI: 10.1002/bimj.201900351

Biometrical Journal

RESEARCH PAPER

Frequentist performances of Bayesian prediction intervals for random-effects meta-analysis

Yuta Hamaguchi[1,2] | Hisashi Noma[3] | Kengo Nagashima[4] | Tomohide Yamada[5] | Toshi A. Furukawa[6]

[1] Department of Statistical Science, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies, Tokyo, Japan

[2] Diagnostics Department, Asahi Kasei Pharma Corporation, Tokyo, Japan

[3] Department of Data Science, The Institute of Statistical Mathematics, Tokyo, Japan

[4] Research Center for Medical and Health Data Science, The Institute of Statistical Mathematics, Tokyo, Japan

[5] Department of Diabetes and Metabolic Diseases, Graduate School of Medicine, University of Tokyo, Tokyo, Japan

[6] Departments of Health Promotion and Human Behavior, Kyoto University Graduate School of Medicine/School of Public Health, Kyoto, Japan

Correspondence
Hisashi Noma, Department of Data Science, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan.
Email: noma@ism.ac.jp

Funding information
Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science, Grant/Award Numbers: JP19H04074, JP17K19808

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

The prediction interval has been increasingly used in meta-analyses as a useful measure for assessing the magnitude of treatment effect and between-studies heterogeneity. In calculations of the prediction interval, although the Higgins–Thompson–Spiegelhalter method is used most often in practice, it might not have adequate coverage probability for the true treatment effect of a future study under realistic situations. An effective alternative candidate is the Bayesian prediction interval, which has also been widely used in general prediction problems. However, these prediction intervals are constructed based on the Bayesian philosophy, and their frequentist validities are only justified by large-sample approximations even if noninformative priors are adopted. There has been no certain evidence that evaluated their frequentist performances under realistic situations of meta-analyses. In this study, we conducted extensive simulation studies to assess the frequentist coverage performances of Bayesian prediction intervals with 11 noninformative prior distributions under general meta-analysis settings. Through these simulation studies, we found that frequentist coverage performances strongly depended on what prior distributions were adopted. In addition, when the number of studies was smaller than 10, there were no prior distributions that retained accurate frequentist coverage properties. We also illustrated these methods via applications to two real meta-analysis datasets. The resultant prediction intervals also differed according to the adopted prior distributions. Inaccurate prediction intervals may provide invalid evidence and misleading conclusions. Thus, if frequentist accuracy is required, Bayesian prediction intervals should be used cautiously in practice.

KEYWORDS
Bayesian prediction, heterogeneity, meta-analysis, prediction interval, random-effects model

1 INTRODUCTION

Random-effect models have been the primary statistical tools for meta-analyses since they enable quantitative evaluation of the average treatment effects accounting for the between-studies heterogeneity (DerSimonian & Laird, 1986; Whitehead & Whitehead, 1991). Conventionally, the "grand mean" parameter has been addressed as a primary estimand of these evidence synthesis studies, but its ability to express a summarized measure of treatment effects among multiple studies is substantially limited because it is solely a point measure that corresponds to the mean of the random-effects distribution. To quantify the true effect and effectiveness in real-world uses of the treatment, more appropriate measures are required that suitably reflect the degree of heterogeneity and magnitude of the treatment effect. The prediction interval was proposed to address this problem; it is defined as an interval that covers the true treatment effect in a future study with certain probability (Higgins, Thompson, & Spiegelhalter, 2009; Riley, Higgins, & Deeks, 2011).

The prediction interval has been gaining prominence in recent meta-analyses because it enables the simultaneous assessment of uncertainties in treatment effects and heterogeneity between studies (IntHout, Ioannidis, Rovers, & Goeman, 2016; Veroniki et al., 2019). The Higgins–Thompson–Spiegelhalter (HTS) method (Higgins et al., 2009) has been most widely used for calculation of the prediction interval. The HTS method is computationally efficient and practically useful, but in a recent study by Partlett and Riley (2017) the HTS prediction interval can have poor coverage properties when the number of studies n is small. The most important reason for this is that the HTS method is based on the large-sample approximation in which n is sufficiently large. In medical meta-analyses, n is usually less than 20 (Kontopantelis, Springate, & Reeves, 2013); therefore, the large-sample approximation can be violated, similar to the problem with inference of the grand mean (Brockwell & Gordon, 2001; Noma, 2011; Veroniki et al., 2019).

An effective alternative approach is Bayesian prediction methods, which have been extensively investigated for general statistical prediction problems (Carlin & Louis, 2009; Gelman et al., 2013). Higgins et al. (2009) also discussed Bayesian approaches in the context of constructing the prediction interval for meta-analyses. Although the Bayesian approach is always "exact" if it is used as a purely subjective method, most cases adopt noninformative prior distributions and use as objective Bayesian approaches (Bodnar, Link, Arendacká, Possolo, & Elster, 2017; Röver, 2020). When using noninformative prior distributions, the resultant predictions (inferences) are good approximations of frequentist predictions (inferences) (Carlin & Louis, 2009; Gelman et al., 2013; Spiegelhalter, Abrams, & Myles, 2004); accordingly, these methods have often been used in practice as if they were frequentist methods. However, a Bayesian prediction interval is rigorously defined as an interval within which a future value falls with a particular subjective probability, and its concordance with the frequentist probability is only assured under large-sample settings (Gelman et al., 2013). In practical situations of meta-analyses in medical research, n is limited (usually <20) (Kontopantelis et al., 2013). Also, several simulation-based studies have indicated that Bayesian inferences are not necessarily accurate in the frequentist sense (Agresti & Min, 2005), and their frequentist performances vary dramatically among prior distributions (Lambert, Sutton, Burton, Abrams, & Jones, 2005). To date, there is no definitive evidence regarding how Bayesian prediction intervals perform in practical meta-analysis situations.

In this article, we conducted extensive simulation studies to assess the frequentist coverage performances of the Bayesian prediction intervals with various 11 noninformative priors. In addition, we provide two illustrative examples of analyses selected from recently published systematic reviews in leading medical journals. Our aim was to obtain certain numerical evidence for the operating characteristics of Bayesian prediction methods, as well as to provide recommendations for the practical use of these methods. We also provide related simulation and real data analysis results for the credible intervals in the Supporting Information.

2 PREDICTION INTERVALS FOR META-ANALYSIS

We consider that there are n clinical trials and y_i $(i = 1, 2, \dots, n)$ is the estimated treatment effect measure in the ith trial. Commonly used effect measures are mean difference, standardized mean difference, risk difference, risk ratio, and odds ratio (Higgins & Green, 2008; Whitehead, 2002); the ratio measures are typically transformed to logarithmic scales. The random-effects model in meta-analyses (DerSimonian & Laird, 1986; Whitehead & Whitehead, 1991) is then defined as

$$\begin{aligned} y_i &\sim N\left(\theta_i, \sigma_i^2\right) \\ \theta_i &\sim N\left(\mu, \tau^2\right), \end{aligned} \tag{1}$$

where θ_i is the true effect size of the ith study and μ is the grand mean parameter. σ_i^2 and τ^2 express within- and across-studies variances; σ_i^2 is usually assumed to be known and fixed to their valid estimates. The across-studies variance τ^2 represents the degree of heterogeneity across studies. Conventionally, the grand mean parameter μ is used as a summary measure of the random-effects meta-analysis as an average treatment effect, and it is estimated as $\hat{\mu} = (\sum_{i=1}^{n} \hat{w}_i y_i)/(\sum_{i=1}^{n} \hat{w}_i)$, where $\hat{w}_i = (\sigma_i^2 + \hat{\tau}^2)^{-1}$; $\hat{\tau}^2$ is an estimator of the heterogeneity variance, for example, the method of moment estimator proposed by DerSimonian and Laird (1986). However, when certain heterogeneity exists, the point measure has a limited ability to utilize meta-analysis evidence for medical policy making and health technology assessments; its effectiveness should be evaluated considering the heterogeneity in the target population. Thus, the prediction interval has been gaining prominence as a way to add useful information involving the uncertainty of the treatment effect and its heterogeneity (IntHout et al., 2016; Veroniki et al., 2019).

The $100(1-\alpha)\%$ prediction interval in meta-analysis is formally defined as an interval that covers the treatment effect θ_{new} in a future study with $100(1-\alpha)\%$ probability (Higgins et al., 2009; Riley et al., 2011). Higgins et al. (2009) proposed a simple plug-in type prediction interval,

$$\left[\hat{\mu} - t_{n-2}^{\alpha}\sqrt{\hat{\tau}_{\text{DL}}^2 + \widehat{\text{Var}}\,[\hat{\mu}]}, \hat{\mu} + t_{n-2}^{\alpha}\sqrt{\hat{\tau}_{\text{DL}}^2 + \widehat{\text{Var}}\,[\hat{\mu}]}\right], \tag{2}$$

where $\hat{\tau}_{\text{DL}}^2$ is the DerSimonian–Laird's method-of-moment estimator (DerSimonian & Laird, 1986) of τ^2, $\widehat{\text{Var}}[\hat{\mu}] = 1/(\sum_{i=1}^{n} (\sigma_i^2 + \hat{\tau}_{\text{DL}}^2)^{-1})$ is the variance estimator of $\hat{\mu}$, and t_{n-2}^{α} is the $100(1-\alpha/2)$ percentile of the t-distribution with $n-2$ degrees of freedom. The HTS prediction interval is based on two approximations: $(\hat{\mu}-\mu)/\sqrt{\widehat{\text{Var}}[\hat{\mu}]}$ is approximately distributed as $N(0,1)$, and $(n-2)(\hat{\tau}_{\text{DL}}^2 + \widehat{\text{Var}}[\hat{\mu}])/(\tau^2 + \widehat{\text{Var}}[\hat{\mu}])$ is approximately distributed as $\chi^2(n-2)$. These approximations are generally not accurate under small- or moderate-n settings. Thus, in the simulation studies of Partlett and Riley (2017) and Nagashima, Noma, and Furukawa (2019b), the coverage probability of the prediction interval for θ_{new} was below the nominal level under general settings of meta-analyses of medical studies, especially when $n < 20$ (Kontopantelis et al., 2013). To address the undercoverage problems, Partlett and Riley (2017) proposed other plug-in–type intervals instead of $\hat{\tau}_{\text{DL}}^2$, for example, the restricted maximum likelihood (REML) estimator. Also, Nagashima et al. (2019b) proposed a parametric bootstrap–based approach using a confidence distribution. In their extensive simulation studies, Nagashima et al. (2019b) showed that the coverage probability of their prediction interval accorded to the nominal level consistently. These prediction intervals are all calculable using the `pimeta` package (Nagashima, Noma, & Furukawa, 2019a) in R (R Foundation for Statistical Computing, Vienna, Austria).

3 | BAYESIAN PREDICTION INTERVALS

3.1 | The Bayesian hierarchical model and prediction

The Bayesian framework using the Markov Chain Monte Carlo (MCMC) is an established prediction method in statistics and represents an alternative effective approach for the prediction problem. Higgins et al. (2009) also discussed the uses of the Bayesian methods in addressing the prediction problem in meta-analyses. In the Bayesian framework, a prediction distribution for the effect θ_{new} is computable by sampling a new study, $\theta_{\text{new}} \sim N(\mu, \tau^2)$ (Higgins et al., 2009; Smith, Spiegelhalter, & Thomas, 1995) by MCMC. A 95% prediction interval for the new study is obtained simply from the 2.5% and 97.5% quantiles of the posterior distribution of θ_{new}.

A Bayesian hierarchical model for the random-effects model (1) is ordinarily constructed assuming prior distributions for the parameters of random-effects distribution μ and τ^2. The purely Bayesian predictions can be conducted under this framework, but here we consider using this Bayesian framework to conduct approximate frequentist predictions using noninformative priors, that is, to provide an accurate prediction interval with a sense of frequentist probability.

3.2 | Prior distributions for approximate frequentist prediction

Various noninformative prior distributions are considered for the meta-analysis model (Lambert et al., 2005). Lambert et al. (2005) also conducted large simulation studies to assess the frequentist performances of the inferences of μ and τ^2

and concluded that they strongly depended on the prior distribution of τ^2. Following their discussions, we considered 11 noninformative priors involving improper priors to assess the impact on prediction accuracy, which is involved in the recently developed `bayesmeta` package in Röver (2020), which uses various modern methods to select a prior distribution. For the grand mean parameter μ, we consistently used a diffuse Gaussian distribution,

$$\mu \sim N(0, 10000). \tag{4}$$

Also, we considered the standard factorable prior distribution, which can be factored into independent marginal, $p(\mu, \tau) = p(\mu) \times p(\tau)$. The 11 prior distributions for τ^2 considered here were as follows.

3.2.1 | Uniform prior distributions for τ

$$p(\tau) \propto 1. \tag{5}$$

This prior distribution is the most intuitive flat noninformative improper prior for the heterogeneity standard deviation parameter τ (Röver, 2020). In addition, as a popular choice in these Bayesian analyses, a proper uniform prior on a limited range can be also considered. One choice would be $\tau \sim Uniform(0, 10)$, which is a flat uniform distribution on $(0, 10)$.

3.2.2 | Uniform prior distribution in $\sqrt{\tau}$

$$p(\tau) \propto \frac{1}{\sqrt{\tau}}. \tag{6}$$

This is the uniform prior in $\sqrt{\tau}$. It has been proposed that a requirement is reasonable for noninformative priors, because it has invariance with respect to rescaling of τ (Jaynes, 1968, 2003). Due to this requirement, a family of improper prior distribution with density is $p(\tau) \propto \tau^a$ $(-\infty < a < \infty)$. As a special case, this prior distribution corresponds to $a = -0.50$ expressing a monotonically decreasing density function; $a = 0$ corresponds to the improper uniform prior in Section 3.2.1.

3.2.3 | Jeffreys prior distribution

$$p(\tau) \propto \sqrt{\sum_{i=1}^{n} \left(\frac{\tau}{\sigma_i^2 + \tau^2} \right)^2}. \tag{7}$$

This is the noninformative Jeffreys prior (Jeffreys, 1946), which is formally given by $p(\mu, \tau) \propto \sqrt{\det(J(\mu, \tau))}$, where $J(\mu, \tau)$ is the Fisher information matrix. In the present Bayesian hierarchical model, the two parameters μ and τ^2 are orthogonal in the sense that the off-diagonal elements of the Fisher information matrix are 0, and the Jeffreys prior corresponds to the Tibshirani's noninformative prior (Tibshirani, 1989); then, the prior density function is given above. This prior also corresponds to the Berger–Bernardo reference prior (Bodnar et al., 2017; Bodnar, Link, & Elster, 2016).

3.2.4 | Berger–Deely prior distribution

$$p(\tau) \propto \prod_{i=1}^{n} \left(\frac{\tau}{\sigma_i^2 + \tau^2} \right)^{1/n}. \tag{8}$$

This is another variation of the Jeffreys prior, provided by Berger and Deely (1988). This prior distribution is also improper and is concordant with the Jeffreys prior when all within-study variances σ_i^2 are equal.

3.2.5 | The proper conventional prior distribution

$$p\left(\tau\right) \propto \prod_{i=1}^{n} \left(\frac{\tau}{\left(\sigma_i^2 + \tau^2\right)^{3/2}} \right)^{1/n}. \tag{9}$$

This is a proper variation of the Jeffreys prior that was proposed by Berger and Deely (1988). This prior distribution is intended as a noninformative, but is used as a proper one for testing or model selection purposes (Berger & Deely, 1988; Berger & Pericchi, 2001).

3.2.6 | DuMouchel prior distribution

$$p\left(\tau\right) = \frac{s_0}{\left(s_0 + \tau\right)^2}, \quad s_0^2 = \frac{n}{\sum_{i=1}^{n} \sigma_i^{-2}}. \tag{10}$$

This is the DuMouchel prior distribution (DuMouchel & Normand, 2000; Spiegelhalter et al., 2004) for the heterogeneity parameter τ. s_0^2 is the harmonic mean of within-study variances σ_i^2. This prior corresponds to a log-logistic distribution for τ that has the mode at 0 and the median at s_0.

3.2.7 | Uniform shrinkage prior distribution

$$p\left(\tau\right) = \frac{2 s_0^2 \tau}{\left(s_0^2 + \tau^2\right)^2}, \quad s_0^2 = \frac{n}{\sum_{i=1}^{n} \sigma_i^{-2}}. \tag{11}$$

This prior distribution is derived as a uniform prior for the average shrinkage factor $S_0(\tau) = s_0^2/(s_0^2 + \tau^2)$ (Daniels, 1999; Spiegelhalter et al., 2004). The median is also s_0, and the forms of the DuMouchel prior and this prior depend on the harmonic mean s_0. The uniform prior for $S_0(\tau)$ is equivalent to a uniform prior of $1 - S_0(\tau) = \tau^2/(s_0^2 + \tau^2)$, which has similar form to the Higgins' I^2 statistic (Higgins & Thompson, 2002).

3.2.8 | Uniform prior distribution in I^2 statistic

$$p\left(\tau\right) = \frac{2\hat{\sigma}^2 \tau}{\left(\hat{\sigma}^2 + \tau^2\right)^2}, \quad \hat{\sigma}^2 = \frac{(n-1)\sum_{i=1}^{n} \sigma_i^{-2}}{\left(\sum_{i=1}^{n} \sigma_i^{-2}\right)^2 - \sum_{i=1}^{n} \sigma_i^{-4}}. \tag{12}$$

This is the uniform prior distribution for Higgins' I^2 statistic (Higgins & Thompson, 2002). As mentioned in Section 3.2.7, this prior density is obtained by substituting the harmonic mean s_0^2 for their average $\hat{\sigma}^2$ from the uniform shrinkage prior; these two priors have similar forms.

3.2.9 | Proper inverse Gamma prior distributions

$$\frac{1}{\tau^2} \sim \mathrm{Gamma}\left(0.001,\ 0.001\right). \tag{13}$$

This prior distribution is the most commonly used semiconjugate prior for the heterogeneity variance parameter (Lambert et al., 2005). The shape of this prior distribution is mostly flat over a wide range, but has a small mode near 0. In addition, we considered a variation of this prior,

$$\frac{1}{\tau^2} \sim \mathrm{Gamma}\left(0.01,\ 0.01\right), \tag{14}$$

TABLE 1 The priors for τ^2 adopted in simulation studies

No.	Prior distributions for τ^2	Characteristics
1	The uniform improper prior in τ (Uniform)	This prior distribution is the most intuitive flat noninformative improper prior for the heterogeneity standard deviation parameter τ.
2	The uniform improper prior in $\sqrt{\tau}$ (Sqrt)	This prior distribution has been proposed that a requirement is reasonable for noninformative priors, because it has invariance with respect to re-scaling of τ.
3	The Jeffreys prior (Jeffreys)	This prior distribution is the noninformative Jeffreys prior, which is formally given by $p(\mu, \tau) \propto \sqrt{\det(J(\mu, \tau))}$, where $J(\mu, \tau)$ is the Fisher information matrix.
4	The Berger–Deely prior (Berger–Deely)	This prior distribution is improper and concordant with the Jeffreys prior when all within-study variances σ_i^2 are equal.
5	The proper conventional prior (Conventional)	This prior distribution is intended as a noninformative, but is used as a proper one for testing or model selection purposes.
6	The DuMouchel prior (DuMouchel)	This prior distribution includes the harmonic mean of within-study variances.
7	The uniform shrinkage prior (Shrinkage)	This prior distribution is derived as a uniform prior for the average shrinkage factor.
8	The uniform prior for I^2 statistic (I2)	This prior distribution is the uniform prior distribution for Higgins' I^2 statistic.
9	The proper uniform prior $U(0, 10)$ (Proper 1)	This prior distribution is a flat uniform distribution as well as Uniform.
10	The proper inverse Gamma prior Gamma$(0, 001, 0.001)$ (Proper 2)	This prior distribution is the most commonly used semiconjugate prior for the heterogeneity variance parameter.
11	The proper inverse Gamma prior Gamma$(0.01, 0.01)$ (Proper 3)	This prior distribution is a vague prior for τ^2, but more informative than the Proper 2.

whose two parameters were changed. This prior is also a vague prior for τ^2, but more informative than the above prior. Using this prior, we can assess the sensitivity of altering the hyperparameters to the frequentist performance of the prediction interval.

4 | SIMULATIONS

4.1 | Designs and settings

We conducted a series of simulation studies to provide certain evidence for frequentist performances of Bayesian prediction intervals for meta-analyses. We adopted 11 priors for τ^2 explained in Section 3. Details of the prior distributions are presented in Table 1. The 95% prediction intervals for a future study were calculated based on the 2.5% and 97.5% quantiles of the posterior distribution of $\theta_{\text{new}} \sim N(\mu, \tau^2)$. For the computations, we used the `bayesmeta` package (Röver, 2020) in R except for Proper 2 and Proper 3 and OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009) for Proper 2 and Proper 3. Also, we added the HTS interval (HTS) as a reference method.

The simulation data were generated mimicking the simulation settings of Brockwell and Gordon (2001, 2007), who consider typical settings of meta-analyses in medical studies that assess an overall odds ratio. The grand mean parameter μ was set to 0, without loss of generality for assessing coverage and precision of the prediction intervals. The within-study variances σ_i^2 were generated from a chi-squared distribution with 1 degree of freedom, multiplied by 0.25, and truncated within an interval [0.009, 0.6]; the mean, median, and 95% probability interval were 0.17, 0.12, and [0.01, 0.55]. We generated the within-trial variances separately within each simulation and used the generated σ_i^2 values when we fitted the priors. The number of studies n and the heterogeneity variance τ^2 were varied for two patterns: (1) n was fixed to 7 or 15, and τ^2 varied among 0.01, 0.02,…, 0.20, and (2) τ^2 was fixed to 0.10 or 0.20, and n varied among 4, 5, 6,…, 20. We used the same simulated trials to the 12 different methods. For each scenario, we simulated 10,000 replications. We assessed the empirical coverage rates of randomly generated $\theta_{\text{new}} \sim N(\mu, \tau^2)$ of the 12 prediction intervals and their empirical expected widths for the 10,000 results. The coverage probabilities are desirable to accord the nominal level, 95%.

We also calculated the 95% credible intervals of the grand mean parameter μ for the 11 priors from the same simulated datasets. We similarly assessed the empirical coverage rates and their empirical expected widths of these credible intervals for μ. We provided these results in e-Appendix A in Supporting Information as supplementary information of these simulation studies.

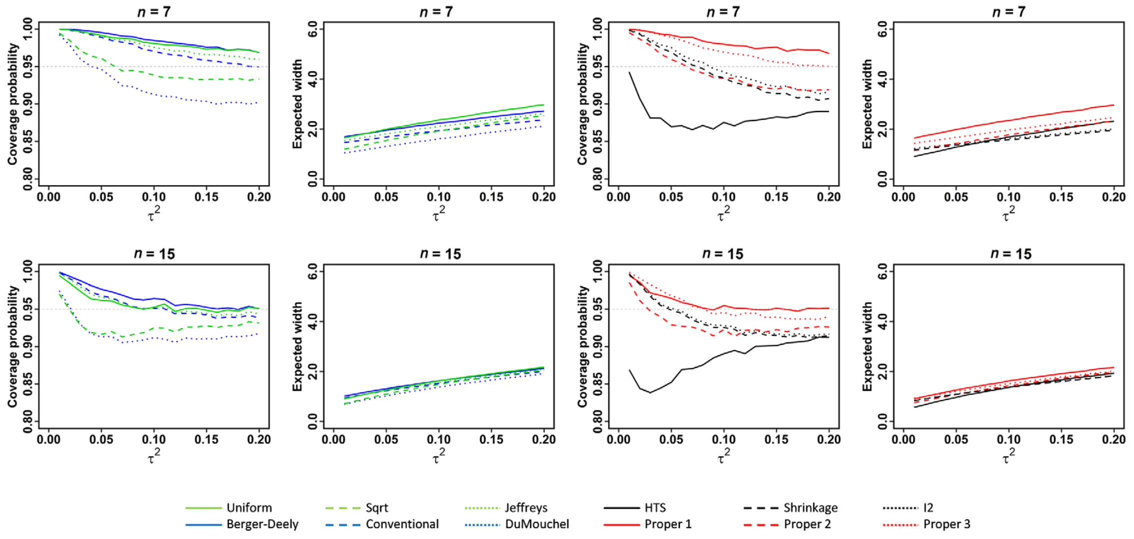

FIGURE 1 Simulation results: coverage probabilities and expected widths of the 95% prediction intervals. Varying τ^2 on 0.01, 0.02, 0.03, …, 0.20 under fixed n (= 7, 15)

4.2 | Results

First, the results of the simulation studies for the scenarios in which n was fixed and τ^2 was varied are presented in Figure 1. Under both settings for the number of studies ($n = 7,\ 15$) settings, most of the Bayesian methods exhibited overcoverage when $\tau^2 = 0.01$; only HTS had coverage probabilities around 0.95. However, when τ^2 grew larger, HTS exhibited undercoverage. HTS consistently exhibited undercoverage when $\tau^2 > 0.01$, and these results were consistent with previous simulation studies (Nagashima et al., 2019b; Partlett & Riley, 2017). Under $n = 7$, the coverage probabilities for Sqrt, DuMouchel, and Proper 2 were below the nominal level when there was certain heterogeneity. Under $n = 15$, they also exhibited undercoverage under smaller τ^2. However, their coverage probabilities were generally larger when τ^2 grew larger. The simulation results of Shrinkage and I2 had similar trends because of the similarities in the shapes of their prior distributions. These two methods also exhibited overcoverage when $\tau^2 < 0.05$; but as heterogeneity got larger, they tended to exhibit undercoverage. For Uniform, Jeffreys, Berger–Deely, Conventional, Proper 1, and Proper 3, the coverage probabilities were above the nominal level regardless of the degree of heterogeneity under $n = 7$; when $\tau^2 = 0.20$, the coverage probabilities of Conventional and Proper 3 were around 0.95. Also, their expected widths were larger than those of the aforementioned priors. Under $n = 15$, they exhibited overcoverage when $\tau^2 < 0.10$, and had accurate coverage probability around the nominal level when $\tau^2 \geq 0.10$, except for Proper 3. Proper 3 exhibited undercoverage when $\tau^2 \geq 0.10$; also, only Berger–Deely tended to exhibit overcoverage. The results of Proper 2 and Proper 3 were quite different, although they were the same parametric inverse gamma priors; the differences indicate sensitivity to the selection of hyperparameters, and suggest that these trends might change if the simulation settings are altered. The overcoverage properties under small τ^2 would be caused by overestimation biases of the heterogeneity variance estimators; some of the simulation results are presented in e-Appendix B. These results reveal that we cannot explicitly specify which priors can provide accurate prediction intervals in general. The expected widths clearly reflected the trends of coverage probabilities. From the above, Uniform, Jeffreys, Berger–Deely, Conventional, and Proper 1 had accurate frequentist coverage properties when $n = 15,\ \tau^2 \geq 0.10$. However, the others exhibited undercoverage or overcoverage and did not provide accurate prediction intervals in the frequentist sense.

Second, the results of simulation studies for the scenarios in which τ^2 was fixed and n was varied are presented in Figure 2. Under $\tau^2 = 0.10$, when the number of studies were extremely small ($n = 4, 5$), DuMouchel, Shrinkage, I2, Proper 2, and HTS had the most accurate coverage probabilities. However, as n got larger, these five methods exhibited undercoverage. In particular, the coverage probabilities of DuMouchel were below 0.90. For Sqrt, the coverage probabilities were above the nominal level when $n \leq 6$, but far below 0.95 when $n > 6$. Uniform, Jeffreys, Berger–Deely, Conventional, Proper 1, and Proper 3 tended to exhibit overcoverage when $n < 12$; when $n \geq 12$, they had accurate coverage probabilities except for Proper 3, which exhibited undercoverage then. Among these priors, Berger–Deely exhibited the greatest degree of overcoverage. Under $\tau^2 = 0.20$, DuMouchel, Shrinkage, I2, Proper 2, and HTS consistently exhibited undercoverage,

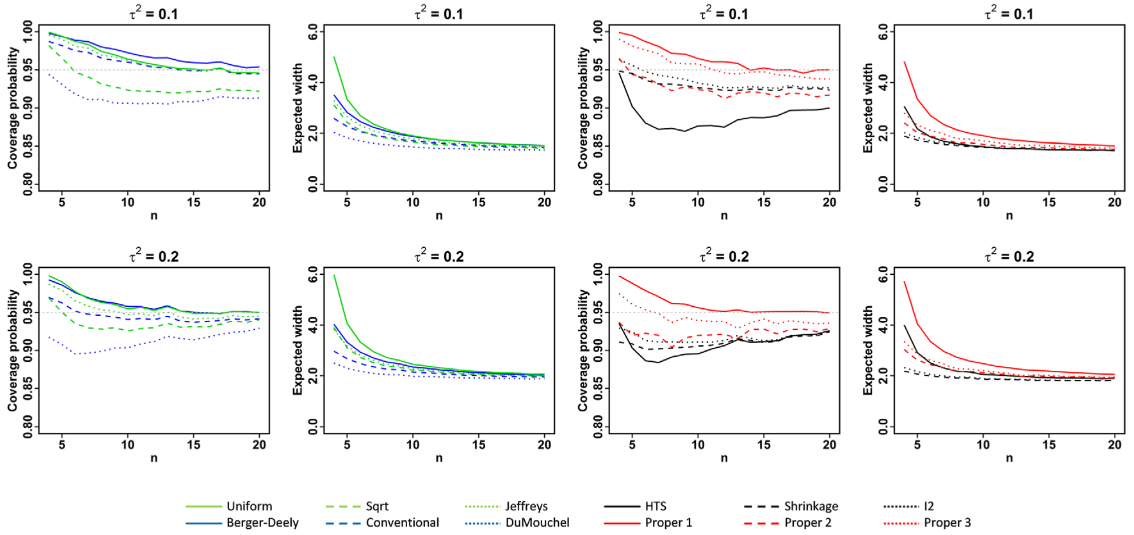

FIGURE 2 Simulation results: coverage probabilities and expected widths of the 95% prediction intervals. Varying n on 4, 5, 6, …, 20 under fixed τ^2 (= 0.1, 0.2)

but when n grew larger, their coverage performances improved. In addition, the coverage probabilities of Sqrt and Conventional were above the nominal level under $n \leq 5$ or 6, but they tended to exhibit undercoverage when $n > 6$, with Sqrt exhibiting undercoverage to a greater degree. Uniform, Jeffreys, Berger–Deely, and Proper 1 exhibited overcoverage under $n \leq 10$, but they had generally accurate coverage probabilities when $n > 10$. Proper 3 exhibited overcoverage when $n \leq 6$, but exhibited undercoverage when n grew larger. In addition, under these settings, the coverage performances of Proper 2 and Proper 3 were also quite different. When n grew larger, the differences got to be smaller, because the relative information of observed data got larger. The expected widths also clearly reflected the trends in coverage probabilities; Uniform and Proper 1 had markedly larger expected widths than the other priors when $n = 4$ or 5. Overall, Uniform, Jeffreys, Berger–Deely, and Proper 1 had accurate frequentist coverage properties when ,$\tau^2 = 0.10, \quad n \geq 12$, and $\tau^2 = 0.20, \quad n \geq 11$; Conventional also achieved nearly accurate coverage performance. However, the others exhibited undercoverage or overcoverage and did not provide accurate prediction intervals in the frequentist sense. Therefore, there are not favorable prediction intervals, especially under $n < 10$, that we can recommend as accurate prediction tools in frequentist sense in meta-analyses of medical studies.

5 | REAL DATA EXAMPLES

For illustrative purposes, we applied Bayesian prediction intervals to two real data examples by Salvo et al. (2016) and Häuser et al. (2009). We present the Bayesian prediction intervals using the 11 reference priors adopted in Section 4. In addition, as reference methods, we also present the HTS interval (HTS), the HTS interval using the Hartung–Knapp variance estimator (HTS–HK), the HTS interval using the Sidik–Jonkman bias-corrected variance estimator (HTS–SJ) (Partlett & Riley, 2017), and the prediction interval–based parametric bootstrap approach using the confidence distribution (pimeta) of Nagashima et al. (2019b) as reference methods. We also present other six real data examples to illustrate the operating characteristics of these methods in e-Appendixes C and D of Supporting Information.

5.1 | DPP-4 data

This meta-analysis was conducted to quantify the risk of hypoglycemia with dipeptidyl peptidase-4 (DPP-4) inhibitors and sulfonylureas compared with placebo and sulfonylureas (Salvo et al., 2016). The outcome was an incidence of hypoglycemia, and the effect measure was risk ratio (RR). Figure 3a presents the prediction intervals of this data. The data included 10 studies and revealed moderate heterogeneity ($\tau^2_{\mathrm{DL}} = 0.02$). The Berger–Deely, Conventional, Jeffreys, Uniform, Proper 1, and Proper 3 intervals were wider than the others. Also, HTS–HK and HTS–SJ were narrower than the

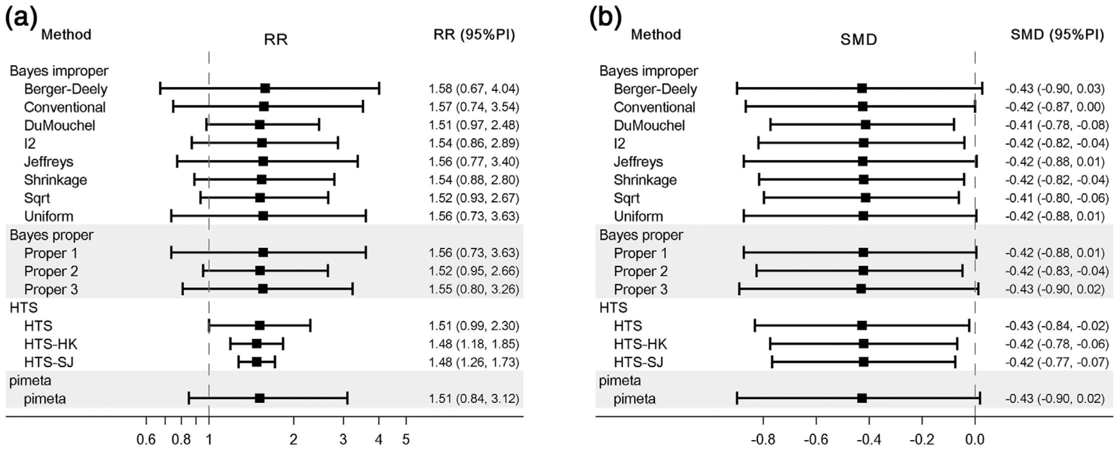

FIGURE 3 The 95% prediction intervals for two illustrative examples: (a) DPP-4 data ($n = 10$) and (b) pain data ($n = 22$)

HTS interval. Although the overall RR was 1.513 (95% CI: [1.219, 1.878]), these prediction intervals might lead different interpretations of the results. The HTS–HK and HTS–SJ intervals did not include 1, and the Bayesian, HTS, and pimeta intervals involved 1.

5.2 | Pain data

Häuser, Bernardy, Üçeyler, and Sommer (2009) conducted a systematic review to compare the treatment effect of antidepressants on pain in patients with fibromyalgia syndrome. The outcomes were the pain questionnaire scores by VAS (visual analog scale), VASFIQ (visual analog scale fibromyalgia impact questionnaire), and NRS (numeric rating scale) and were synthesized as standardized mean difference (SMD). Figure 3b presents the prediction intervals. The data included 22 studies, and the heterogeneity variance estimate was 0.03 ($P = .012$). Among the Bayesian prediction intervals, the Berger–Deely, Conventional, Jeffreys, Uniform, Proper 1, and Proper 3 intervals included 0; on the other hand, the DuMouchel, I2, Shrinkage, Sqrt, and Proper 2 intervals did not. As for the frequentist methods, the HTS, HTS–HK, and HTS–SJ intervals did not include 0, but the pimeta interval did. The interpretations might differ based on the priors we adopted for this meta-analysis.

6 | DISCUSSION

The prediction interval has been gaining prominence in recent meta-analyses and could become a standard statistical output in meta-analyses in the near future because of its effectiveness for the assessment of heterogeneity and uncertainties of treatment effects in target populations (Higgins et al., 2009; IntHout et al., 2016; Riley et al., 2011; Veroniki et al., 2019). Bayesian prediction methods represent a useful approach in practices, but our study revealed that Bayesian prediction intervals are not necessarily accurate in the frequentist sense under practical situations.

Especially, when the number of studies n is smaller than 10, our simulation studies showed that the Bayesian prediction intervals based on all 11 reference priors did not have favorable coverage performances. Also, the coverage performances were quite different. Thus, we must choose the prior distributions carefully. In particular, some priors exhibited serious undercoverage properties (Sqrt, DuMouchel, Shrinkage, I2, Proper 2, and Proper 3) and might underestimate the heterogeneity and uncertainty in practice. Besides, Uniform, Jeffreys, Berger–Deely, Conventional, and Proper 1 exhibited good coverage performances, in particular when $n > 10$; however, they tended to provide overly wide prediction intervals when $n \leq 10$ and might yield vague evidence.

If prediction intervals are too vague or too narrow, they can directly influence the conclusions of systematic reviews, providing misleading evidence for health technology assessments and policy making. Therefore, Bayesian prediction intervals should be carefully used in practice, and if there exist other accurate alternatives, they should be recommended. Certainly, it is not problematic if researchers wish to conduct purely Bayesian prediction with subjective probability, but most Bayesian applications in meta-analyses are conducted as objective Bayesian frameworks (Lambert et al., 2005; Röver, 2020). Similar discussions have been provided in relation to the undercoverage problem of standard confidence intervals of grand mean parameter μ, and various improved methods have been developed (Brockwell & Gordon, 2001; Noma, 2011; Veroniki et al., 2019). Besides, for prediction intervals, rich methods do not yet exist. Currently, the confidence distribution approach of Nagashima et al. (2019a, 2019b) represents a suitable choice for accurate predictions; in their simulation-based numerical evaluations, they exhibited accurate coverage properties in general. In this regard, development of more rich methods for accurate predictions is needed and represents an important priority for future work in research synthesis methodology.

There would be several further issues in our study. First, we assumed the within-study variances are known in the simulations, but they are usually estimated from observed datasets. It possibly provides additional uncertainty to the overall results. However, in this study, we addressed the primary purpose of our simulations to evaluate the operational performances of the prediction intervals under the conditions "the model assumptions were completely correct." Actually, we confirmed the Bayesian prediction intervals could not perform well even if the assumptions were correct. Besides, their influences were empirically evaluated in previous simulation studies (Noma, Nagashima, Kato, Teramukai, & Furukawa, 2020; Sidik & Jonkman, 2007; Sugasawa & Noma, 2019), and they were not so influential to the overall results under broad settings. Second, we used the conventional DerSimonian–Laird estimate for computing the HTS prediction interval. Although there are several sophisticated estimators such as the REML estimator for the heterogeneity variance (Sidik & Jonkman, 2007), currently the DerSimonian–Laird estimator is mostly adopted for the HTS prediction interval in practices. We used this conventional estimator as a current standard method, but alternative effective approaches would be commonly used in future; several popular frequentist approaches were evaluated by simulation studies in Partlett and Riley (2017). Third, we adopted the conventional DerSimonian–Laird-type normal–normal model in this study, other sophisticated methods have been developed in recent studies; for example, effective approaches for rare event settings (Böhning, Mylona, & Kimber, 2015) and random-effects Poisson regression model (Beisemann, Doebler, & Holling, 2020). Further investigations for the prediction intervals based on these sophisticated methods would be needed, but we guess similar results would be obtained for these alternative models. Also, the normality assumption for the random-effects models were critically discussed in recent studies (Jackson & White, 2018; Noma et al., 2020). Further methodological developments to exceed the restrictions for conventional modeling strategies would also be expected.

In conclusion, inaccurate prediction intervals may provide invalid evidence and misleading conclusions. Our simulation-based evidence would indicate that the Bayesian prediction intervals should be used cautiously in practice, if frequentist accuracy is required.

ACKNOWLEDGMENTS

This study was supported by Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (Grant numbers: JP19H04074, JP17K19808).

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

The meta-analysis datasets used in Section 5 are parts of the published data from Salvo et al. (2016) and Häuser et al. (2009).

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Hisashi Noma https://orcid.org/0000-0002-2520-9949
Kengo Nagashima https://orcid.org/0000-0003-4529-9045

REFERENCES

Agresti, A., & Min, Y. (2005). Frequentist performance of Bayesian confidence intervals for comparing proportions in 2 × 2 contingency tables. *Biometrics*, *61*(2), 515–523.

Beisemann, M., Doebler, P., & Holling, H. (2020). Comparison of random-effects meta-analysis models for the relative risk in the case of rare events: A simulation study. *Biometrical Journal*. *62*(7), 1597–1630.

Berger, J. O., & Deely, J. (1988). A Bayesian approach to ranking and selection of related means with alternatives to analysis-of-variance methodology. *Journal of the American Statistical Association*, *83*(402), 364–373.

Berger, J. O., & Pericchi, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison. In P. Lahiri (Ed.), *Model section, volume 38 of IMS Lecture Notes* (pp. 135–193). Beachwood, OH: Institute of Mathematical Statistics.

Bodnar, O., Link, A., Arendacká, B., Possolo, A., & Elster, C. (2017). Bayesian estimation in random effects meta-analysis using a non-informative prior. *Statistics in Medicine*, *36*(2), 378–399.

Bodnar, O., Link, A., & Elster, C. (2016). Objective Bayesian inference for a generalized marginal random effects model. *Bayesian Analysis*, *11*(1), 25–45.

Böhning, D., Mylona, K., & Kimber, A. (2015). Meta-analysis of clinical trials with rare events. *Biometrical Journal*, *57*(4), 633–648.

Brockwell, S. E., & Gordon, I. R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine*, *20*(6), 825–840.

Brockwell, S. E., & Gordon, I. R. (2007). A simple method for inference on an overall effect in meta-analysis. *Statistics in Medicine*, *26*(25), 4531–4543.

Carlin, B. P., & Louis, T. A. (2009). *Bayesian methods for data analysis*. New York, NY: Chapman & Hall.

Daniels, M. (1999). A prior for the variance in hierarchical models. *Canadian Journal of Statistics*, *27*(3), 567–578.

DerSimonian, R., & Laird, N. M. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188.

DuMouchel, W. H., & Normand, S. L. (2000). Computer modeling strategies for meta-analysis. In D. K. Stangl & D. A. Berry (Eds.), *Meta-analysis in medicine and health policy* (pp. 127–178). Boca Raton, FL: CRC Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Häuser, W., Bernardy, K., Üçeyler, N., & Sommer, C. (2009). Treatment of fibromyalgia syndrome with antidepressants: A meta-analysis. *JAMA*, *301*(2), 198–209.

Higgins, J. P. T., & Green, S. (2008). *Cochrane handbook for systematic reviews of interventions*. Chichester, UK: Wiley-Blackwell.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558.

Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society, Series A*, *172*(1), 137–159.

IntHout, J., Ioannidis, J. P., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open*, *6*(7), e010247.

Jackson, D., & White, I. R. (2018). When should meta-analysis avoid making hidden normality assumptions? *Biometrical Journal*, *60*(6), 1040–1058.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, *4*(3), 227–241.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London, Series A*, *196*(1007), 453–461.

Kontopantelis, E., Springate, D. A., & Reeves, D. (2013). A re-analysis of the Cochrane Library data: The dangers of unobserved heterogeneity in meta-analyses. *PloS One*, *8*(7), e69930.

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., & Jones, D. R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, *24*(15), 2401–2428.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*(25), 3049–3067.

Nagashima, K., Noma, H., & Furukawa, T. A. (2019a). pimeta: Prediction intervals for random-effects meta-analysis. R package version 1.1.3. https://cran.r-project.org/web/packages/pimeta/.

Nagashima, K., Noma, H., & Furukawa, T. A. (2019b). Prediction intervals for random-effects meta-analysis: A confidence distribution approach. *Statistical Methods in Medical Research*, *28*(6), 1689–1702.

Noma, H. (2011). Confidence intervals for a random-effects meta-analysis based on Bartlett-type corrections. *Statistics in Medicine*, *30*(28), 3304–3312.

Noma, H., Nagashima, K., Kato, S., Teramukai, S., & Furukawa, T. A. (2020). Flexible random-effects distribution models for meta-analysis. *arXiv*, 2003.04598.

Partlett, C., & Riley, R. D. (2017). Random effects meta-analysis: Coverage performance of 95% confidence and prediction intervals following REML estimation. *Statistics in Medicine*, *36*(2), 301–317.

Riley, R. D., Higgins, J. P. T., & Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *British Medical Journal*, *342*, d549.

Röver, C. (2020). Bayesian random-effects meta-analysis using the bayesmeta R package. *Journal of Statistical Software*, *93*(6), 1–51.

Salvo, F., Moore, N., Arnaud, M., Robinson, P., Raschi, E., De Ponti, F., … Pariente, A. (2016). Addition of dipeptidyl peptidase-4 inhibitors to sulphonylureas and risk of hypoglycaemia: Systematic review and meta-analysis. *British Medical Journal*, *353*, i2231.

Sidik, K., & Jonkman, J. N. (2007). A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine*, *26*(9), 1964–1981.

Smith, T. C., Spiegelhalter, D. J., & Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine*, *14*(24), 2685–2699.

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*. Chichester, UK: John Wiley & Sons.

Sugasawa, S., & Noma, H. (2019). A unified method for improved inference in random-effects meta-analysis. *Biostatistics*, kxz020. https://doi.org/10.1093/biostatistics/kxz020.

Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika*, *76*(3), 604–608.

Veroniki, A. A., Jackson, D., Bender, R., Kuss, O., Langan, D., Higgins, J. P. T., … Salanti, G. (2019). Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis. *Research Synthesis Methods*, *10*(1), 23–43.

Whitehead, A. (2002). *Meta-analysis of controlled clinical trials*. Chichester, UK: Wiley.

Whitehead, A., & Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomised clinical trials. *Statistics in Medicine*, *10*(11), 1665–1677.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Hamaguchi Y, Noma H, Nagashima K, Yamada T, Furukawa TA. Frequentist performances of Bayesian prediction intervals for random-effects meta-analysis. *Biometrical Journal*. 2021;63:394–405. https://doi.org/10.1002/bimj.201900351