Received: 17 March 2023 | Revised: 4 October 2023 | Accepted: 3 November 2023

DOI: 10.1002/jrsm.1685

RESEARCH ARTICLE

Research Synthesis Methods WILEY

Bayesian random-effects meta-analysis with empirical heterogeneity priors for application in health technology assessment with very few studies

Jona Lilienthal[1] | Sibylle Sturtz[1] | Christoph Schürmann[1] | Matthias Maiworm[1,2] | Christian Röver[3] | Tim Friede[3] | Ralf Bender[1]

[1]Department of Medical Biometry, Institute for Quality and Efficiency in Health Care (IQWiG), Köln, Germany

[2]SALETELLIGENCE GmbH, Bielefeld, Germany

[3]Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

Correspondence
Jona Lilienthal, Department of Medical Biometry, Institute for Quality and Efficiency in Health Care (IQWiG), Köln, Germany.
Email: jona.lilienthal@iqwig.de

Abstract

In Bayesian random-effects meta-analysis, the use of weakly informative prior distributions is of particular benefit in cases where only a few studies are included, a situation often encountered in health technology assessment (HTA). Suggestions for empirical prior distributions are available in the literature but it is unknown whether these are adequate in the context of HTA. Therefore, a database of all relevant meta-analyses conducted by the Institute for Quality and Efficiency in Health Care (IQWiG, Germany) was constructed to derive empirical prior distributions for the heterogeneity parameter suitable for HTA. Previously, an extension to the normal-normal hierarchical model had been suggested for this purpose. For different effect measures, this extended model was applied on the database to conservatively derive a prior distribution for the heterogeneity parameter. Comparison of a Bayesian approach using the derived priors with IQWiG's current standard approach for evidence synthesis shows favorable properties. Therefore, these prior distributions are recommended for future meta-analyses in HTA settings and could be embedded into the IQWiG evidence synthesis approach in the case of very few studies.

KEYWORDS

external information, heterogeneity, hierarchical model, meta-analysis, prior distribution

Highlights

What is already known

- Random-effects meta-analysis with very few studies is frequently unreliable because of uncertainties in the estimation of the between-study heterogeneity.
- Bayesian methods can be used in such situations but require specification of appropriate prior distributions.

- Whereas uninformative priors are commonly used for the effects, there is some debate regarding appropriate choices of the prior for the heterogeneity parameter.

What is new

- Data extracted from reports of IQWiG, a German HTA agency, are used to derive informative prior distributions suitable for applications in HTA. Differences to the current procedure used by IQWiG are investigated.

Potential impact for research synthesis methods readers

- A possible evidence synthesis process is described that combines the proposed Bayesian approach in the current IQWiG procedure.

1 | INTRODUCTION

Health technology assessment (HTA) reports often include results from multiple studies, necessitating a systematic overview. If possible, a meta-analysis is performed to combine the individual study results into an interpretable treatment effect estimate. In most cases, a random-effects model is preferred due to heterogeneity between the studies that is generally to be expected. When there are only very few studies available (less than five), the application of a random-effects model can be difficult because the heterogeneity between studies cannot be reliably estimated. In such situations, the current approach used by the Institute for Quality and Efficiency in Health Care (IQWiG) is rather complex and involves the calculation and comparison of multiple meta-analytic models (see also Section 4.1).[1]

Even though IQWiG's General Methods[1] state that Bayesian approaches may "also be an option," they are not regularly used for meta-analyses. More widespread application of Bayesian methods might simplify analyses compared to the current, rather intricate procedure. While the choice of a uniform (uninformative) prior distribution for the treatment effect parameter is usually uncontroversial, the specification of an appropriate prior for the between-study heterogeneity is often more complex.[2] In case of many studies, a uniform prior may be appropriate as well, but for the common case of very few studies, using a proper weakly informative prior is often advantageous or even necessary.[2,3] Informative prior distributions for the between-study heterogeneity may be based on empirical information from previous research, and various approaches for summarizing empirical information for the heterogeneity parameter are available.[4–6] Röver et al.[6] proposed an extended normal-normal hierarchical model for Bayesian random-effects meta-analysis and primarily discussed methodological aspects in detail. In the present investigation, we apply this method to data from IQWiG reports to derive prior distributions for the heterogeneity parameter. We also recommend useful priors for IQWiG reports and compare the results of Bayesian analysis using these priors with IQWiG's current evidence synthesis procedure.

It has been most relevant for us that the prior distributions are tailored to and applicable to IQWiG's HTA purposes. Furthermore, we would like to take particular care to derive conservative specifications. Considering that *underestimation* of heterogeneity is potentially more harmful, while *overestimation* may be seen as a conservative form of bias, we would rather prefer prior specifications favoring slightly larger heterogeneity values.

For this purpose, we set up a database of relevant meta-analyses conducted in IQWiG projects. Its composition and a descriptive summary are given in Section 2. Section 3 briefly introduces the extended normal–normal hierarchical Bayesian model proposed by Röver et al.[6] used to derive informative priors. This model is then applied to the IQWiG data set, the results are presented, and recommendations are made. In Section 4, we elaborate on IQWiG's current approach for evidence synthesis with only a few studies and compare it to the results of Bayesian meta-analyses using the recommended prior distributions. The manuscript concludes with a discussion of the most important findings and results (Section 5).

2 | DATABASE OF META-ANALYSES

2.1 | Set up of database

All reports published by IQWiG until December 31, 2021, were screened for meta-analyses, both benefit and dossier assessments. Data from forest plots presented in the reports were extracted and entered into a spreadsheet. Meta-analyses done by IQWiG itself as well as by third parties were included. We omitted meta-analyses that were only referred to in the text (without a figure). It was irrelevant whether the analysis itself was of direct interest or whether it was performed only to test the suitability of individual pairwise comparisons (for example, to check a homogeneity assumption in the context of network meta-analyses). Forest plots of sensitivity and specificity were not considered, neither were sensitivity analyses or subgroup analyses.

Forest plots for binary outcomes presenting no events in either arm for all but one study were excluded, as the pooled odds ratio (OR) or relative risk (RR) reduces to the effect of the single study in that case. If both responder analyses (binary data) and evaluations based on a mean difference (MD) were available for operationalization of an endpoint, both evaluations were included in the database and subsequent analyses. If meta-analyses based on both MD and standardized MD (SMD) were available, only the results based on SMD were considered. In various projects, analyses at different time points were presented in forest plots each on their own. This could be due to different evaluation times within studies or to different data cut-offs within studies. We considered all forest plots as relevant for the database, regardless of which was the most relevant in a given report. Studies that were repeatedly included in assessments, like due to follow-up data later submitted for addenda, were also included multiple times in the database. This also applied to individual arms of multi-arm studies, which could be included in different objectives or even reports. We did not adjust our analyses for such multiple uses of the same data for pragmatic reasons.

2.2 | Descriptive analysis

We screened a total of 867 IQWiG reports, 134 of which contained at least 1 meta-analysis. A total of 919 analyses used binary data, 645 continuous data, and 112 time-to-event analyses. Two-thirds of the analyses compared pharmacological interventions (against other pharmacological interventions or placebo), 32% compared non-pharmacological interventions (against other non-pharmacological interventions or placebo, and only 1% compared pharmacological against non-pharmacological interventions. Most analyses (46%) investigated outcomes of morbidity, 32% adverse events, 10% health-related quality of life, 7% mortality, and 6% were others such as biological markers or surrogates. Tables A1, A2, and A3 in Appendix A in Data S1 provide numbers as on different data types along with a comparison of endpoint categories and intervention types.

Some studies feature multiple times in the data set, if included in more than one report, but this is rarely the case: The study total was 1075 and 69 of these were included in two different reports. However, we did not check for studies being included under different names (study name or publication name). A greater limitation of our data is that studies are included with each relevant endpoint and therefore may be included multiple times. The median number of endpoints included per study and report in the full data set is 3 (IQR 2–6), with ∼10% of studies having more than 10 endpoints included. However, different effect measures were extracted from these studies. Therefore, subsetting on a certain effect measure, these studies appear less often than in the full data set.

In case of binary data presented in 2×2 contingency tables, we calculated both OR and RR, yielding 883 meta-analyses with OR as the effect measure and 917 meta-analyses with RR. A few reports did not report contingency tables but aggregated effect measures (point estimate and standard error) only, so these numbers differ slightly between OR and RR. Another 112 meta-analyses were based on the hazard ratio (HR) and 645 meta-analyses on the SMD. In total, data from 2557 meta-analyses were available for the eventual analysis. In this section, we present detailed results using OR as the effect measure of interest. Descriptive analyses for other effect measures are provided in Appendix A in Data S1.

For the OR, just under half, that is, 431 out of 883 analyses (49%), considered exactly 2 studies, and 679 (77%)

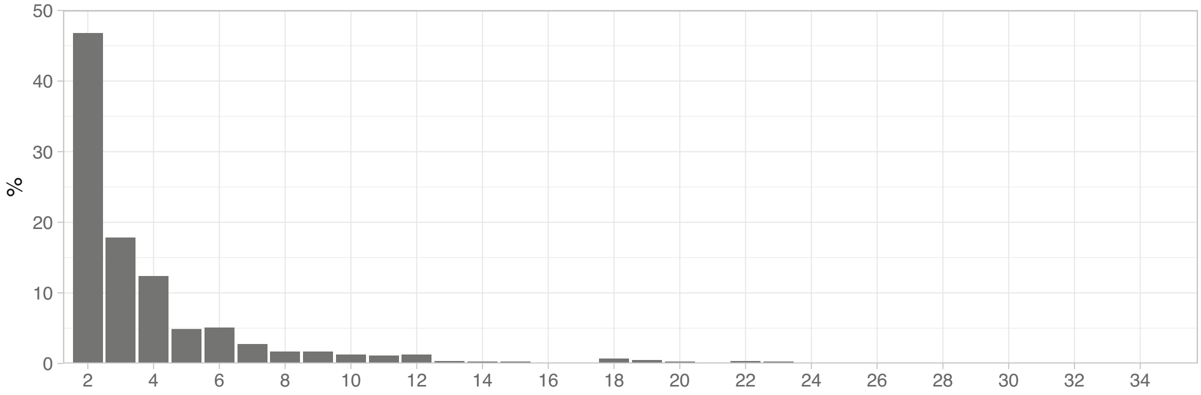

FIGURE 1 Bar chart of the number of studies included in each meta-analysis. Effect measure: OR.

	N	Min	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	Max
Number of MAs per report	100 reports	1	3	5.5	10	63
Number of studies per MA	883 MAs	2	2	3	4	35
Sample size per study	3569 studies	20	558	1019	1879	229,588

TABLE 1 Descriptive statistics on numbers of meta-analyses (MAs) per report, numbers of studies per MA, and sample sizes. Effect measure: OR.

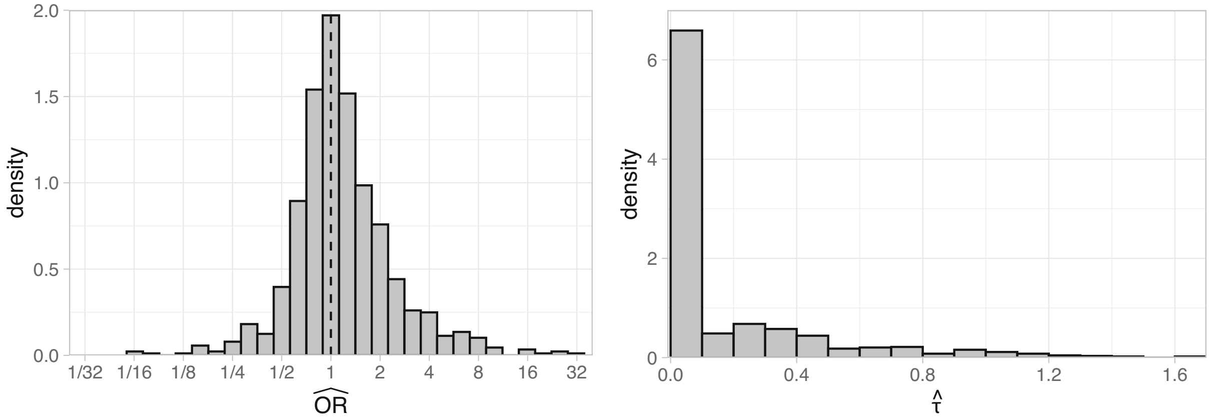

FIGURE 2 Left: histogram of the effect estimates; dashed line marks null-effect. Right: histogram of heterogeneity estimates ($\hat{\tau}$). Effect measure: OR.

considered fewer than 5 studies (Figure 1). Descriptive data for the number of meta-analyses per report, the number of studies per meta-analysis, and their respective sample size are given in Table 1. Some IQWiG reports address more than one research question. As these naturally consist of more analyses, reports with multiple research questions are counted separately for each question in this table.

Heterogeneous results that would not be pooled according to IQWiG standard operating procedures ($p \leq 0.05$ by the heterogeneity test) were present in 66 meta-analyses (7%). Among these, 19 meta-analyses were of 2 studies, 15 of 3 studies, and 10 of 4 studies. Regardless of the heterogeneity and the number of studies, the results are pooled by using a random-effects meta-analysis according to the Knapp-Hartung method with Paule-Mandel estimator for the heterogeneity parameter to give an idea of the distribution of the heterogeneity τ.[7] A histogram of the estimated effects is shown in Figure 2 (left panel). Point estimates of the OR range from 0.059 to 28.888, their median is 1.063. A total of 693 meta-analyses (78%) resulted in a statistically significant overall effect ($p \leq 0.05$).

Figure 2 (right panel) presents a histogram of estimated τ from all 883 meta-analyses. In 561 (64%) of these, τ was estimated as zero. These estimates of zero were based on 261 meta-analyses with 2 studies, 105 with 3 studies, 68 with 4 studies, 31 with 5 studies, and 96 with more than 5 studies. The proportion of non-zero heterogeneity estimates (36% for all meta-analyses) remains at a similarly high level if the pool of meta-analyses is restricted to those with at least 5 studies (38%) or those that are not based on heterogeneous study results (31%) (see Table A5 in Appendix A in Data S1 for the results for all effect measures).

3 | DERIVATION OF INFORMATIVE PRIORS FOR THE HETEROGENEITY PARAMETER

In this section, empirical information extracted from IQWiG reports (see Section 2) is used to derive an informative prior distribution for application in future Bayesian random-effects meta-analyses. Multiple meta-analyses are accommodated in a joint model to facilitate inference on the heterogeneity parameter(s). This approach has been described in the literature for binary and continuous outcomes[4,5] and has recently been extended.[6] In such a Bayesian framework, the posterior predictive distribution of the heterogeneity parameter constitutes a prior distribution for future meta-analyses. In this manuscript, the methodology will be reported only in condensed form, for more information as well as an example code of an implementation using JAGS, see Röver et al.[6]

3.1 | Framework

Consider the framework of the normal-normal hierarchical model (NNHM) for meta-analysis, which here is extended to accommodate data from several meta-analyses (indexed by j, with $j = 1,...,N$):

$$y_{ij} \mid \mu_j, \tau_j \sim \text{Normal}\left(\mu_j, \sigma_{ij}^2 + \tau_j^2\right), \tag{1}$$
$$\mu_j \mid \mu_P, \sigma_P \sim \text{Normal}\left(\mu_\text{p}, \sigma_\text{p}^2\right), \tag{2}$$

in which the ith treatment effect estimate of the jth meta-analysis, y_{ij}, is normally distributed with expectation μ_j and variance $\sigma_{ij}^2 + \tau_j^2$ (where $i = 1,...,k_j$). The expectations μ_j are assigned a vague normal hyperprior with parameters μ_p and σ_p to effectively stratify by meta-analysis. The variance $\sigma_{ij}^2 + \tau_j^2$ consists of the study-specific uncertainty of estimation, σ_{ij}^2, and the heterogeneity parameter τ_j^2 which describes the inter-study variability in the jth meta-analysis. While σ_{ij} is generally assumed to be known, the heterogeneity τ_j is often hard to assess, but is of primary interest here. In the next hierarchy level, the "population" of heterogeneity parameters is modeled via a parametric distribution P, and using a hyperprior H for its parameters:

$$\tau_j \mid \theta \sim P(\theta), \qquad \theta \sim H. \tag{3}$$

Several distribution families for modeling the heterogeneity parameters are used and compared in the present manuscript, covering different shapes and tail behaviors:

- Half-normal distribution: $\tau_j \mid s \sim \text{HN}(s), \quad s \sim \text{Unif}(0,b)$,
- Exponential distribution: $\tau_j \mid s \sim \text{Exp}(1/s), \quad s \sim \text{Unif}(0,b)$,
- Log-normal distribution: $\tau_j \mid s,t \sim \text{LN}(\log(s),t)$, $s \sim \text{Unif}(0,b), \quad t \sim \text{Unif}(0,b)$,
- Half-logistic distribution: $\tau_j \mid s \sim \text{HL}(s), \quad s \sim \text{Unif}(0,b)$.

In all cases, s denotes a *scale parameter*, and we set $b = 10$ as a large upper bound for its uniform hyperparameter distribution (given that we are focusing on ORs here). The log-normal distribution in addition possesses a *shape parameter* t, for which we similarly specify a uniform prior.

Inference for these models is facilitated using Markov chain Monte Carlo (MCMC) methods. We are primarily interested in two figures: the heterogeneity distribution's parameter(s) θ (usually: its scale s, and possibly also its shape parameter t), as well as the posterior predictive distribution of a "new" heterogeneity value $\tau^\star$, both of which can be derived directly from the MCMC samples. For practical use in a Bayesian meta-analysis model, we want to obtain a more manageable parametric distribution to adequately describe the posterior predictive distribution. Different options are possible for this step[6]:

1. Calculate a point estimate of the parameter's distribution, $\widehat{\theta}$, and use the conditional distribution $p\left(\tau^\star \mid \widehat{\theta}\right)$.
2. Take the parameter's uncertainty into account and calculate a mixture distribution.
3. Approximate the sample distribution of $\tau^\star$ by fitting a suitable distribution using, for example, maximum likelihood estimation or method of moments.

In the following application to our data set, we will present the MCMC sample distribution of both θ and $\tau^\star$ and compare the different options for obtaining a parametric distribution.

3.2 | Application to IQWiG data

The IQWiG data set consists of meta-analyses based on ORs, RRs, HRs, and SMDs. The described procedure is employed separately for each effect measure. We used the statistical software R and JAGS for computation[8–10] (JAGS code for the model is available in the online supplement of Röver et al.[6]). This section reports the results for the effect measure OR in detail and briefly summarizes the results for RRs, HRs, and SMDs. Comparisons between the results of the effect measures and their consequences for the recommendations regarding prior distributions are also presented.

3.2.1 | Odds ratios

Figure 3 illustrates the predictive distributions (of $\tau^\star$) resulting from using different distribution families modeling the conditional distribution $\tau_j \mid s$ (see Equation (3)). The distributions have a median lying between 0.08 (LN) and 0.11 (HN) and an interquartile range between 0.10 (LN) and 0.14 (HN). The range between the lower and upper quartiles spans slightly higher values for the half-normal model than for the others, but overall the results appear consistent in terms of the general dimensions of the distributions.

Figure 4 shows histograms of the posterior distribution of the half-normal scale parameter s, as well as the corresponding predictive distribution of $\tau^\star$ resulting from a half-normal model for τ_j. The scale parameter's posterior lies mainly between 0.14 and 0.20 (2.5% and 97.5% quantiles). Two parametric approximations of the predictive distribution are illustrated in the right panel. The half-normal distribution using the posterior mean of the scale parameter, $\bar{s} = 0.167$, is shown in blue. To account for the scale parameter's uncertainty, we also fitted a half-Student-t distribution as an alternative (yellow line).[6] As apparent in the figure, both approximations are very similar and only differ slightly. This is also evident from the large degrees-of-freedom parameter of

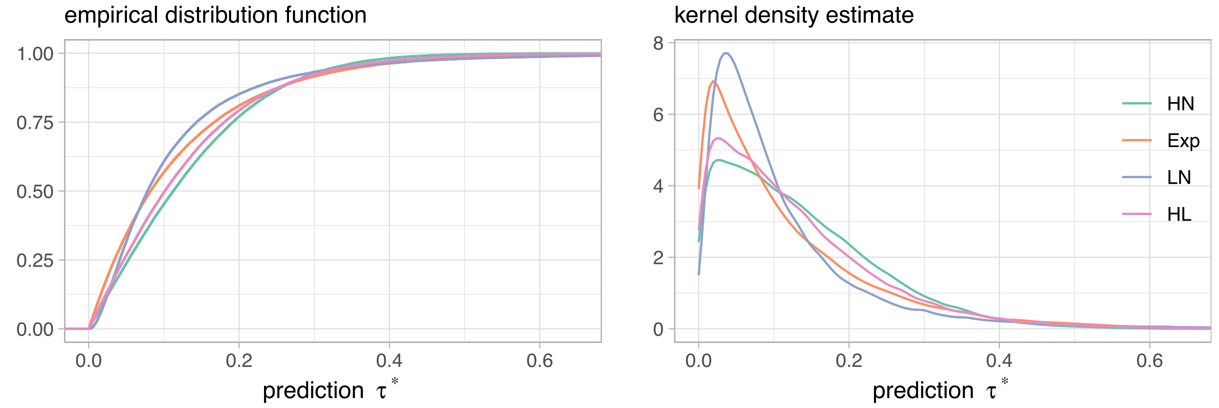

FIGURE 3 Empirical distribution functions (left) and kernel density estimates (right) of the predictive distribution of $\tau^\star$ resulting from assuming different distribution families for τ_j. Effect measure: OR.

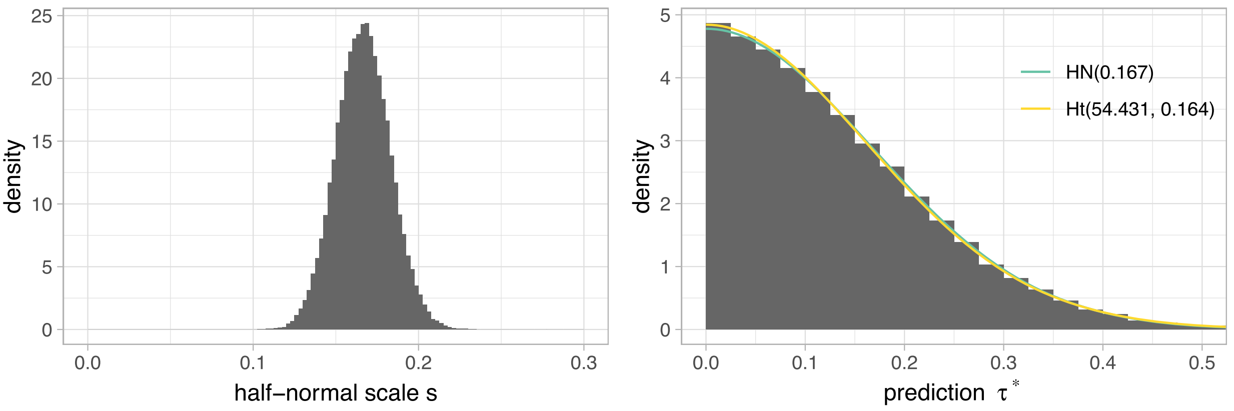

FIGURE 4 Results for the half-normal model. Left panel: posterior distribution of the half-normal scale parameter s; right panel: predictive distribution of $\tau^\star$, along with half-normal and half-Student-t fits. Effect measure: OR.

the half-Student-t distribution, $\nu = 54.4$. Both approximations seem to represent the predictive distribution in an adequate way. Therefore, with simplicity in mind, one might choose the half-normal distribution with scale $s = 0.167$ as a suitable prior distribution.

Similarly, the results using the exponential, log-normal, and half-logistic distribution as a model for τ_j were analyzed. In each case, the point estimate approach led to sufficient approximations of the respective predictive distributions. The resulting fitted distributions are compared in Figure 5. Variations of the inherent distribution shapes, placing differing emphasis on smaller or higher values of the heterogeneity parameter, can be seen.

In the remaining part of our work we focus on the half-normal as a model for τ_j for several reasons: We found that the different models seem to make little difference regarding the predictive distribution of $\tau^\star$. The half-normal prior offers the advantage of being a simple model with only one scale parameter. Its short upper tail prevents too extreme heterogeneity values while not placing too much probability mass on near-zero values compared to the other distributions. This avoids an increased risk of under-coverage due to an increased risk of underestimation of the between-study variability. Furthermore, the half-normal is frequently used and has been investigated in extensive simulation studies.[11] In addition, the effect of the shape of different prior distributions on the result of a Bayesian meta-analysis was demonstrated in Röver et al. (2021).[2] For a set of different heterogeneity priors with a common median, very little sensitivity to the prior's distribution family was observed.

3.2.2 | Other effect measures

We also performed analogous analyses for HR, RR, and SMD data (see Appendix B in Data S1 for the

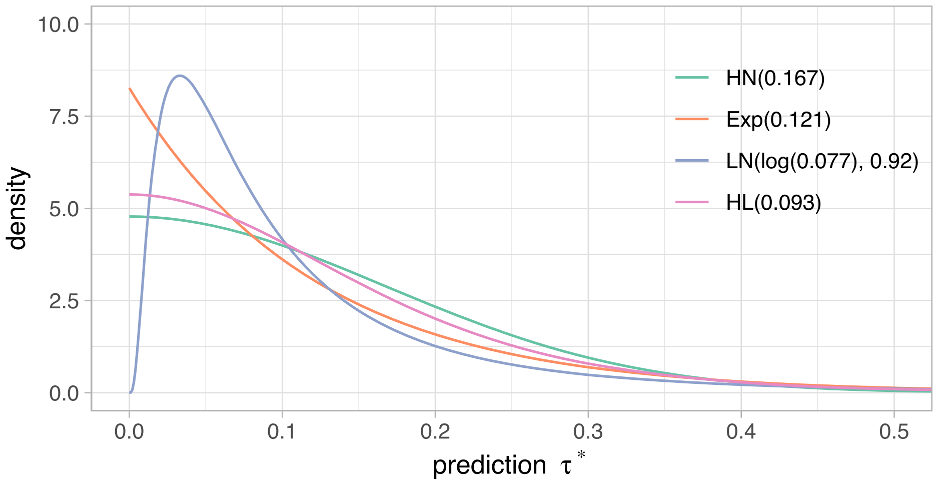

FIGURE 5 Densities of point estimate approximations of the predictive distributions in models using the corresponding distribution families as the heterogeneity prior. Effect measure: OR.

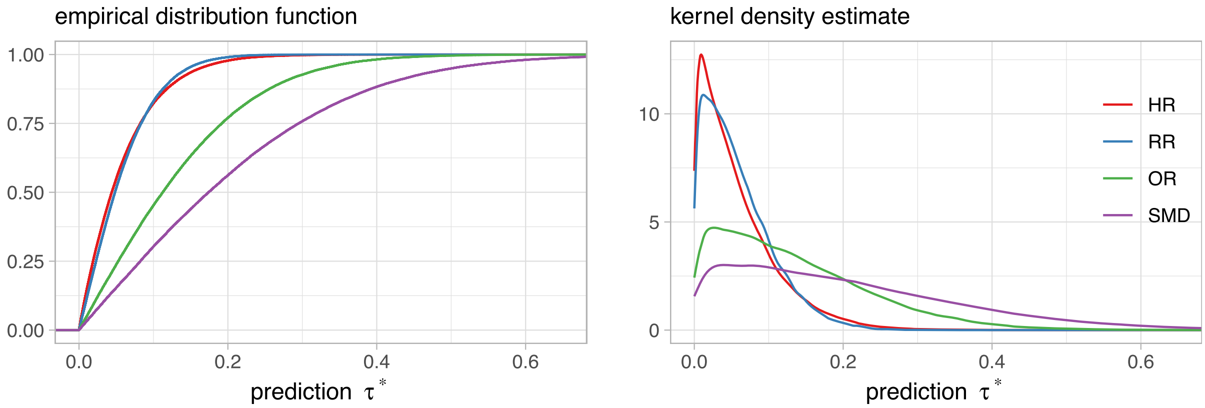

FIGURE 6 Empirical distribution functions (left) and kernel density estimates (right) of the predictive distribution of τ^* resulting using the half-normal distribution for different effect measures.

corresponding figures). Figure 6 illustrates the predictive distributions in those models when using the half-normal model for τ_j. Compared with the results for OR, the distribution is shifted to higher values in case of SMD data. HR and RR data both resulted in distributions having more emphasis on smaller values. Another observation is that the results using HR and RR data were very similar.

Figure 7 illustrates the half-normal scale parameters' posterior distributions side-by-side for the different effect measures. Different color shades and the given numerical values indicate the 50%, 90%, and 99% quantiles. The same shift as was seen in the previous figure is apparent here: the posterior of the scale parameter has higher values for SMD data and lower values for HR and RR data. The differing distribution widths between HR and RR data reflect different uncertainties about the parameter estimates resulting from the different numbers of meta-analyses available for the effect measures (645, 883, 917, and 112 meta-analyses for SMD, OR, RR, and HR, respectively).

3.2.3 | Sensitivity analyses

To analyze the robustness of our results presented in the previous sections, we split the data set of reports reporting meta-analyses into several parts using the year of commission of the report ("old" reports before 2010 vs. "new" reports after 2014), the type of the report (drug assessments vs. others), as well as a combination of both. Summary statistics regarding endpoint categories and intervention types of the analyses of those subsets are included in Tables A1 and A2 in Appendix A in Data S1. Then, the same analyses were calculated for each subset and the results were compared. There is a notable correlation between the two factors: dossier assessments, a special type of report for early benefit assessment of drugs, started in 2011 and was the most prevalent drug assessment from thereon. After 2014, they account for half of the reports containing meta-analyses. The main difference we found is in the type of the report. In comparison to other reports, like those on non-pharmacological interventions, dossier assessments

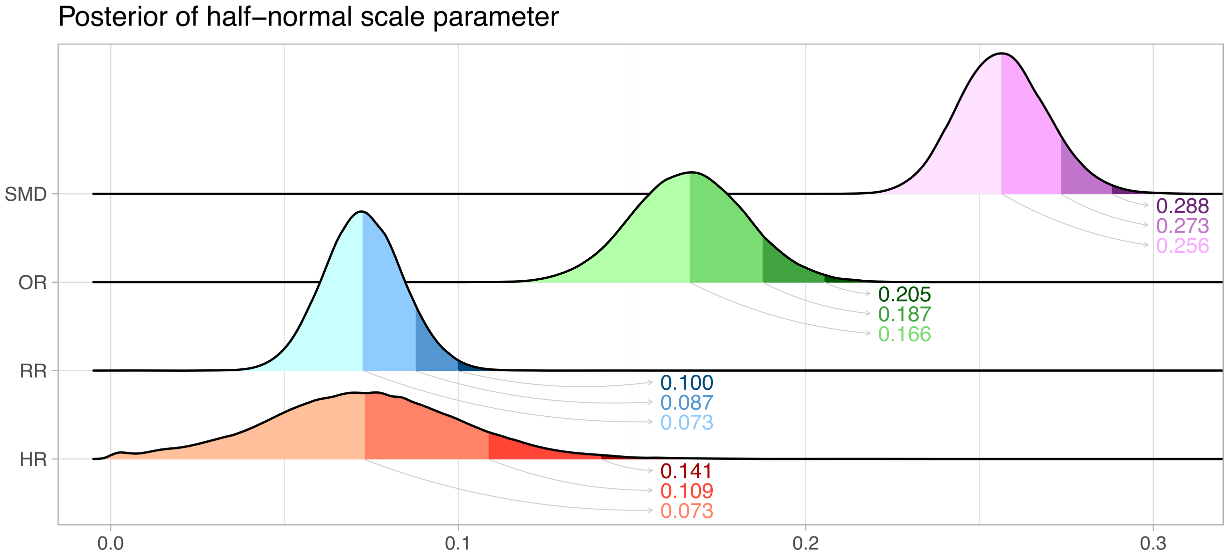

FIGURE 7 Posterior distributions of the half-normal scale parameter for different effect measures. Different color shades and the given numerical values indicate the 50%, 90%, and 99% quantiles, respectively.

tend to generally consist of fewer and more similar studies (rarely more than two). Therefore, restricting the data set to meta-analyses of drug assessments results in a predictive distribution of the heterogeneity parameter that is shifted toward zero. Differences between older and newer reports are smaller when the interaction with the type of report is taken into account. Overall, to account for these differences and to avoid underestimation of heterogeneity, the prior distribution should be chosen in a way that is not overly focused on smaller heterogeneity values.

3.2.4 | Recommendations

The analyses of IQWiG data presented above shall now be translated into a set of readily applicable, empirically motivated recommendations. An important aspect is that one may generally consider larger heterogeneity as a more conservative assumption; while zero heterogeneity means that the analysis simplifies to a *common-effect* approach, larger heterogeneity implies less certainty in the overall estimate, as well as less sharing of information between the included studies. This results in wider confidence intervals of the common effect and therefore fewer statistically significant results. Specification of a *stochastically larger* heterogeneity prior (or a larger prior scale parameter) will therefore usually be a more conservative assumption.[2,12] Given that half-normal distributions are fairly common and established as prior distributions for the heterogeneity parameter, and since in the present application they also tended to yield the largest prior medians, we choose to use the half-normal model among the alternatives investigated here, which yielded similar results anyway. To safeguard against potential underestimation of the heterogeneity, we will also derive recommendations with a view of the upper tails of the scale parameters' posterior distributions.

On the basis of Figure 7 we suggest the use of

1. HN(0.1) for the effect measures RR and HR
2. HN(0.2) for the effect measure OR
3. HN(0.3) for the effect measure SMD

as prior distributions for future Bayesian random-effects meta-analyses. We used approximately the 95% quantiles of the scale parameters' posteriors, and rounded these to the nearest value with a single decimal place. This also has a practical reason: By not choosing a precise value we ensure that updates to our data set will not result in immediate changes to our recommendations. However, in Section 4.1 we will compare results from using the median values and our recommendations.

It is important to note that these suggested priors are not intended for general meta-analysis application but are specifically tailored for use in analyses of HTA investigations implemented in a similar fashion to IQWiG reports. While one should not naively assume exchangeability and immediately apply these priors in different contexts, the inferred magnitudes of heterogeneity will certainly constitute important and relevant signposts and will be helpful in the discussion and justification of related prior specifications. In the following section, we will investigate the consequences of using these recommendations compared to the current procedure used in IQWiG reports.

4 | COMPARISON WITH IQWIG'S CURRENT APPROACH

This section first describes the current evidence synthesis approach used by IQWiG in more detail.[6] In the next step, the results of IQWiG's current approach are compared to the application of Bayesian meta-analyses using the recommended priors.

4.1 | IQWiG's current evidence synthesis approach

The first step is the decision of whether a common-effect model (frequently also called fixed-effect model) is justifiable or not. If there are only two studies, homogeneity is regularly assumed and a common-effect model is calculated, unless there is substantial doubt about the homogeneity. This pragmatic approach is chosen to avoid the extreme model-change step from one to two studies, which results from the very uncertain estimation of heterogeneity in this case. If a common-effect model is not justifiable (especially if three or more studies are available), it is assessed whether a *meaningful* overall effect can be estimated by application of a random-effects model. For this, overall effect estimates are calculated using the Knapp-Hartung method with and without ad-hoc variance correction (abbreviated as KH and KH-VC, respectively)[13,14] as well as using the method of DerSimonian-Laird (DSL), a normal approximation based, for example, on the Paule-Mandel heterogeneity estimate.[15] By comparing the confidence intervals' widths, either KH-VC (if the confidence interval of KH is narrower than that of DSL) or KH is chosen. The chosen method's confidence interval is then compared with the confidence intervals of the individual studies and is considered *informative* if it is fully contained in the union of the individual intervals (since a narrower interval reflects greater confidence in the location of the true effect). An informative overall effect is called *meaningful* if the conclusion on statistical significance is in agreement with that according to DSL. A decision is derived as follows:

1. If the overall effect is considered to be "meaningful," the estimation according to KH (either with or without variance correction) is used.
2. If the overall effect is considered "not meaningful," a qualitative summary of study results (QSSR) is conducted. A QSSR allows drawing a conclusion without calculating a pooled estimate. The conclusion is based either on the prediction interval or on conditions regarding the number of studies with effect estimates in the same direction as well as the number of studies with statistically significant results.[1] However, since no pooled estimate is calculated, a possible benefit is non-quantifiable.

4.2 | Comparison between the Bayesian and the current IQWiG approach

According to our recommendations from Section 3, we performed Bayesian analyses using HN(0.1)-priors for meta-analyses of HR and RR endpoints, HN(0.2)-priors for ORs, and HN(0.3)-priors for the analyses based on SMDs. For each of these, the analysis yields a point estimate along with a 95% credible interval. We consider an effect as *statistically significant* if its credible interval excludes the value of the null effect (i.e., 1 in the case of ratios and 0 in the case of SMD). For comparison, we also derived frequentist random-effects point and interval estimates according to DerSimonian and Laird, based on the Paule-Mandel heterogeneity estimate.[15,16]

We present overall results for the complete set of meta-analyses (see Appendix C in Data S1 for subgroups depending on the number of studies including 2, 3, 4, 5 or more, and 2–4 studies, with the latter representing the case of "very few studies"). The conclusion according to IQWiG's current methods[1,17] is taken as the reference value. IQWiG's conclusion is merely a specific algorithm based on comparing the results of various models and including a QSSR (see previous Section). It comprises a few rater-based judgments regarding model suitability and meta-analytic study weights that we replaced by automated decision rules to make the whole procedure computable.

Taking the conclusion by IQWiG's current algorithm as a reference in the evaluation of the results is not intended to establish IQWiG's approach as truth or a gold standard. However, as the underlying "true" effect is unknown, a useful reference standard is given by the current approach. Consequently, an effect as by the Bayesian approach is considered a "false positive" if the internal algorithm results in "no effect." Likewise, evidence of an effect is considered a "false negative" if the Bayesian approach results in an inconclusive effect but the internal algorithm results in "evidence of an effect." The remaining possible combinations are all regarded as concordant, that is, both procedures result in either an effect or no effect. Additionally, we also consider the subset of analyses consisting of statistically significant effects in all included studies. Within this subset, the number of false negative cases is assessed. From a decisioner's perspective, this is an unambiguous situation and any meta-analytic procedure should result in sufficient evidence of an effect as it would be counter-intuitive to remain unsure if all studies show significant results. Especially in this case, we expect concordance to be perfect and would

TABLE 2 Proportions of disagreement regarding statistical significance between IQWiG's current and the Bayesian approach using a half-normal prior (HN_{prop}: with our proposed parameters 0.1 (RR, HR), 0.2 (OR) and 0.3 (SMD), respectively; HN_{med}: with the median value of the parameter's posterior) or random-effects analysis as by DerSimonian and Laird (DSL) instead. Database is all IQWiG reports.

Effect			% disagreeing		
Measure	**Set of meta-analyses**	**# analyses**	**HN_{prop}**	**HN_{med}**	**DSL**
RR	All	917	10	9	8
	No sufficient evidence of effect as by IQWiG procedure	666	9	9	7
	Sufficient evidence of effect as by IQWiG procedure	251	13	9	9
	All studies statistically significant	64	3	3	3
HR	All	112	4	4	3
	No sufficient evidence of effect as by IQWiG procedure	80	3	3	3
	Sufficient evidence of effect as by IQWiG procedure	32	9	9	3
	All studies statistically significant	14	0	0	0
OR	All	883	10	10	8
	No sufficient evidence of effect as by IQWiG procedure	639	9	9	7
	Sufficient evidence of effect as by IQWiG procedure	244	16	15	10
	All studies are statistically significant	63	3	3	5
SMD	All	645	18	17	8
	No sufficient evidence of effect as by IQWiG procedure	428	8	10	9
	Sufficient evidence of effect as by IQWiG procedure	217	36	32	7
	All studies are statistically significant	81	12	10	0

tolerate only a minimal discrepancy. Note that there is no rule for cases with opposing effect directions, for instance, a positive effect resulting from one procedure and a negative effect from the other procedure. A check of the data set revealed that such a situation did not occur. The proportions of concordant, false negative and false positive effects are presented in Table 2.

While we see fairly high proportions of overall agreement between the Bayesian and the current IQWiG approach for all of the effect measures, the more important statistics are the proportions of false positives and false negatives: False positive proportions range between 3% and 9% which seems quite acceptable. However, false negative proportions reach higher levels ranging from 9% to 36%. While a proportion smaller than 10% might be acceptable, proportions exceeding 20% seem too high to be of use in practice. In the special case of only statistically significant studies included, the false negative proportions reach more acceptable levels between 3% and 12% but which are too high, nevertheless.

As a comparison, we have also included the results of Bayesian meta-analyses if not the half-normal distribution with the previously recommended scales is used but the half-normal with a scale corresponding to the median value of the parameter's posterior distribution. Using those smaller scale parameters leads to smaller heterogeneity values in each specific analysis and therefore to shorter confidence intervals. Hence, the rate of "false negative" is reduced while the rate of "false positives" is increased (see Table 2 and Appendix C). However, the differences are not substantial.

The high proportions of false negative results contradict using the Bayesian approach as a sole and universal method. It seems sensible to combine this approach with a qualitative summary of the study results in those cases in which the Bayesian approach does not result in sufficient evidence of an effect but a qualitative summary of study results does. In this case, the qualitative summary of the study results is regarded as more reliable than the Bayesian approach and would override any conclusions from it in these cases. However, as a consequence, the effect size would not be able to be determined. In IQWiG terminology, the extent of added benefit would be "non-quantifiable."[18] Therefore, incorporating the Bayesian approach into the evidence synthesis would more often result in a quantifiable effect than with the previous approach.

5 | DISCUSSION AND PERSPECTIVES

In IQWiG reports, evidence synthesis is used if results from multiple studies are available for one or more endpoints. In general, frequentist methods are currently used for meta-analyses. While the assumption of a common-effect model is often hardly justifiable, random-effects meta-analyses are frequently unreliable in the case of very few studies (less than five). Due to high uncertainty in the estimated heterogeneity parameter, the resulting confidence intervals for the treatment effect may end up being too wide or too narrow. In a complex process, different methods (KH with and without variance correction, DSL) are applied and their results are compared to assess if the overall treatment effect estimate is considered "meaningful" or if a QSSR should be conducted.[1]

Bayesian random-effects meta-analysis can be a useful alternative in this situation to simplify that process by replacing the involved calculation and comparison of different models with a single meta-analysis based on a specified prior distribution. In this context, the application of Bayesian meta-analyses is able to properly account for uncertainty in the heterogeneity parameter. The result can be seen as a compromise between the overly conservative estimation using KH, that often even leads to implausibly wide, uninformative confidence intervals, and the too-liberal estimation of DSL or the common-effect model in the case of true heterogeneity. By specifying a reasonable prior distribution for the heterogeneity, both extremes, the frequently occurring estimates of zero and implausibly high heterogeneity estimates, are effectively avoided. The aim of our work was to derive informative prior distributions to be applied in Bayesian random-effects meta-analyses building on the results from earlier IQWiG reports and to compare the resulting inferences with those based on the IQWiG's standard procedure. In situations with very few studies, in which the heterogeneity cannot be estimated reliably, a valid estimation of a treatment effect is facilitated using a Bayesian random-effects meta-analysis as introduced within this paper.

We propose the combined application of Bayesian meta-analysis with the QSSR to achieve congruent decisions. In comparison to the sole application of QSSR, the combined approach avoids the outcome of a "non-quantifiable" result. The suggested new approach for very few studies in situations in which pooling generally seems meaningful (i.e., no statistically significant heterogeneity test) is described below:

1. For two studies the common-effect meta-analysis is calculated, unless strong reasons indicate otherwise. In analyses of meta-analyses with only two studies (see Appendix C in Data S1), we found that Bayesian meta-analysis offers no advantages over the current approach. For pragmatic reasons, we therefore continue with the current approach in these cases.
2. In cases of 3 or 4 studies (and for 2 studies if common-effect is clearly inappropriate), a Bayesian random-effects meta-analysis using the proposed prior distributions (HN(0.1) for HRs and RRs, HN(0.2) for ORs and HN(0.3) for SMDs) is calculated and compared to the result of the QSSR. If only QSSR yields evidence in favor of an effect, this holds, although the effect cannot be quantified. If both QSSR and the Bayesian approach suggest evidence in favor of an effect, the Bayesian results are used to quantify the treatment effect.

So, for the case of four or fewer studies, evidence synthesis would be based on either the common effect model or a Bayesian meta-analysis combined with QSSR. Frequentist random-effects models would only be applied in case of five or more studies. But, if the new approach appears to yield useful results, it could possibly also be applied to cases of 5 or more (or less than 3) studies in the future. However, little difference between KH and the Bayesian approach is to be expected in the case of many studies included. Using such a combined approach would guarantee some consistency to prior assessments while reducing the problem of non-quantifiable effects while enhancing the simplicity and rigor of the assessment procedure. In the situation of 3 or 4 studies, IQWiG's current approach claims evidence in 172 meta-analyses in our data set. In 40 (23%) of these, the decision is made using QSSR; therefore no quantification is possible. Using the proposed approach of combining QSSR with the Bayesian approach, evidence of an effect is determined in 155 meta-analyses. Only in 9 (6%) of these analyses it is not possible to quantify the treatment effect because the result of the Bayesian analysis is not in agreement with QSSR. In addition to reducing the rate of unquantifiable effects in cases where evidence is claimed, a pooled effect estimate may also be useful in the absence of evidence for an effect.

Our work is similar to Turner et al.,[5] where prior distributions for binary outcomes are derived using 14,886 meta-analyses within the Cochrane Database of Systematic Reviews (CDSR). Similarly, for continuous outcomes, Rhodes et al.[4] analyzed 6492 meta-analyses within the CDSR. The authors differentiated the between-study heterogeneity distributions for 80 different settings of outcome type (e.g., mortality, quality of life/functioning, adverse events) and intervention comparison type (e.g., pharmacological vs. placebo/control). Our analyses differ from this in important ways. We did not distinguish between outcome types and intervention comparison types. Our goal was to obtain general prior distributions that could be used for analysis with the aspects of generality and simplicity in mind. A more complex set of possible distributions depending on outcome and intervention types and probably even more categories would contradict this approach. More practical problems are that our (already limited) database would be drastically decreased under such constraints. Because some endpoints may be classified as either morbidity or adverse events, it is beneficial if the outcome of the meta-analysis does not depend on the specific choice of an a priori distribution. However, differences between intervention comparison types were briefly investigated, but no clear tendencies were apparent. Therefore, we decided to use the data without splitting it up and rounding up the parameters of the prior distributions.

Both Turner et al. and Rhodes et al. also gave prior distributions for all meta-analyses without restriction to outcome or intervention type. For ORs, Turner et al.[5] proposed $\tau^2 \sim \mathrm{LN}(-2.56, 1.74^2)$, and for SMDs, Rhodes et al.[4] proposed $\log(\tau^2) \sim t_5(-3.44, 2.59^2)$. Both distributions have more distributional weight at higher values of heterogeneity than our proposed distributions. In Table B1 in Appendix C in Data S1 a comparison of summary statistics on the untransformed scale of τ is given. For ORs the prior of Turner et al. is generally shifted to higher values, for SMDs a clear difference only appears in the upper quartile.

The reason for this is that our analysis is based on a rather specific data set of IQWiG's HTA reports. The PICO framework is more restrictive than in the more general Cochrane reviews. The selected studies are therefore more similar, resulting in heterogeneity distributions shifted toward zero. As a result, our recommendations are not general and it is unclear whether or to what extent these might be transferable to other applications. However, they are likely suitable for applications of other HTA agencies with similar HTA questions as IQWiG. In the light of the forthcoming implementation of the EU HTA regulation, in which assessments of health technologies will be conducted on a joint European level, it is important that empirical priors suitable for HTA applications are available.

A limitation is the multiple uses of individual study results in the reports and the inclusion of multiple meta-analyses of the same endpoint at different time points. Since we did not correct for multiplicity here, it might be that some of the presented inferences appear more certain than appropriate. Furthermore, the suggested approach is compared only with IQWiG's current approach, we did no simulations to evaluate its performance in general. However, we considered this sufficient for the present purposes, as both approaches lead to very similar solutions.

In summary, new empirical prior distributions for the heterogeneity parameter are derived, which allowed us to include Bayesian random-effects meta-analysis in the evidence synthesis approach in the situation of very few studies for application in HTA. The Bayesian approach seems to be a useful compromise between the liberal DerSimonian-Laird method and the conservative Knapp-Hartung method and allows the quantification of the extent of added benefit more frequently than frequentist and QSSR methods alone.

AUTHOR CONTRIBUTIONS

Jona Lilienthal: Writing – original draft; writing – review and editing; conceptualization; methodology; formal analysis. **Sibylle Sturtz:** Writing – original draft; writing – review and editing; conceptualization; methodology; formal analysis; data curation. **Christoph Schürmann:** Writing – original draft; writing – review and editing; conceptualization; methodology; formal analysis. **Matthias Maiworm:** Writing – review and editing; conceptualization; methodology; formal analysis. **Christian Röver:** Writing – review and editing; conceptualization; methodology. **Tim Friede:** Writing – review and editing; conceptualization; supervision. **Ralf Bender:** Writing – review and editing; conceptualization; supervision.

ACKNOWLEDGEMENTS

We would like to thank the reviewers for their helpful comments, which have greatly improved this work. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflict of interest.

ORCID

Jona Lilienthal https://orcid.org/0000-0002-3745-5370
Christian Röver https://orcid.org/0000-0002-6911-698X
Tim Friede https://orcid.org/0000-0001-5347-7441
Ralf Bender https://orcid.org/0000-0002-2422-4362

REFERENCES

1. Institute for Quality and Efficiency in Health Care. General Methods: Version 6.1. 2022 https://www.iqwig.de/methoden/general-methods_version-6-1.pdf
2. Röver C, Bender R, Dias S, et al. On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. *Res Synth Methods*. 2021;12(4):448-474. doi:10.1002/jrsm.1475
3. Bender R, Friede T, Koch A, et al. Methods for evidence synthesis in the case of very few studies. *Res Synth Methods*. 2018;9(3):382-392. doi:10.1002/jrsm.1297
4. Rhodes KM, Turner RM, Higgins JP. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *J Clin Epidemiol*. 2015;68(1):52-60. doi:10.1016/j.jclinepi.2014.08.012
5. Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JP. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Stat Med*. 2015;34(6):984-998. doi:10.1002/sim.6381

6. Röver C, Sturtz S, Lilienthal J, Bender R, Friede T. Summarizing empirical information on between-study heterogeneity for Bayesian random-effects meta-analysis. *Stat Med*. 2023;42(14): 2439-2454. doi:10.1002/sim.9731
7. Veroniki AA, Jackson D, Bender R, et al. Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis. *Res Synth Methods*. 2019;10(1):23-43. doi:10.1002/jrsm.1319
8. R Core Team. R: A Language and Environment for Statistical Computing. 2022 https://www.R-project.org/
9. Plummer M. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. *Proc 3rd Int Workshop Distrib Stat Comput*. 2003;124(125.10):1-10.
10. Plummer M. rjags: Bayesian Graphical Models using MCMC. 2022 https://CRAN.R-project.org/package=rjags
11. Friede T, Röver C, Wandel S, Neuenschwander B. Meta-analysis of few small studies in orphan diseases. *Res Synth Methods*. 2017;8(1):79-91.
12. Röver C. Bayesian random-effects meta-analysis using the Bayesmeta R package. *J Stat Softw*. 2020;93(6):1-51. doi:10.18637/jss.v093.i06
13. Hartung J. An alternative method for meta-analysis. *Biom J*. 1999;41(8):901-916. doi:10.1002/(Sici)1521-4036(199912)41:83.0.Co;2-W
14. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Stat Med*. 2003;22(17):2693-2710. doi:10.1002/sim.1482
15. DerSimonian R, Laird NM. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177-188. doi:10.1016/0197-2456(86)90046-2
16. Higgins JPT, Thomas J, Chandler J, et al., eds. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd ed. Wiley & Sons; 2019.
17. Schulz A, Schürmann C, Skipka G, Bender R. Performing meta-analyses with very few studies. In: Evangelou E, Veroniki AA, eds. *Meta-Research: Methods and Protocols. 2345 of Methods in Molecular Biology*. Humana Press; 2022:91-102.
18. Skipka G, Wieseler B, Kaiser T, et al. Methodological approach to determine minor, considerable, and major treatment effects in the early benefit assessment of new drugs. *Biom J*. 2016; 58(1):43-58. doi:10.1002/bimj.201300274

AUTHOR BIOGRAPHIES

Jona Lilienthal is Research Associate at the Department of Medical Biometry at IQWiG.

Sibylle Sturtz is Research Associate at the Department of Medical Biometry at IQWiG.

Christoph Schürmann was Research Associate at the Department of Medical Biometry at IQWiG.

Matthias Maiworm was Research Associate at the Department of Medical Biometry at IQWiG.

Christian Röver is Research Associate at the Department of Medical Statistics, University Medical Center Göttingen. He is a member of several professional societies including the Society for Research Synthesis Methodology. He serves as Associate Editor for Research Synthesis Methods.

Tim Friede is Head of the Department of Medical Statistics and Professor of Biostatistics at the University Medical Center Göttingen. He serves as Associate Editor for several journals including Statistics in Medicine. He is a member of several professional societies including the Society for Research Synthesis Methodology.

Ralf Bender is Head of the Department of Medical Biometry at IQWiG and Adjunct Professor at the University of Cologne. He is Academic Editor of PLOS ONE and member of several professional societies including the Society for Research Synthesis Methodology.

SUPPORTING INFORMATION

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

How to cite this article: Lilienthal J, Sturtz S, Schürmann C, et al. Bayesian random-effects meta-analysis with empirical heterogeneity priors for application in health technology assessment with very few studies. *Res Syn Meth*. 2024;15(2):275-287. doi:10.1002/jrsm.1685