RESEARCH ARTICLE

Bayesian model-averaged meta-analysis in medicine

František Bartoš[1] | **Quentin F. Gronau**[1] | **Bram Timmers**[1] |
Willem M. Otte[2,3] | **Alexander Ly**[1,4] | **Eric-Jan Wagenmakers**[1]

[1]Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands

[2]Department of Pediatric Neurology, UMC Utrecht Brain Center, University Medical Center Utrecht and Utrecht University, Utrecht, The Netherlands

[3]Biomedical MR Imaging and Spectroscopy Group, Center for Image Sciences, University Medical Center Utrecht and Utrecht University, Utrecht, The Netherlands

[4]Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

Correspondence
Eric-Jan Wagenmakers, Department of Psychology, University of Amsterdam, PO Box 15906, 1001 NK Amsterdam, The Netherlands.
Email: ej.wagenmakers@gmail.com

Funding information
Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Numbers: 016.Vici.170.083, 406.16.528, NWA.1228.191.045

We outline a Bayesian model-averaged (BMA) meta-analysis for standardized mean differences in order to quantify evidence for both treatment effectiveness δ and across-study heterogeneity τ. We construct four competing models by orthogonally combining two present-absent assumptions, one for the treatment effect and one for across-study heterogeneity. To inform the choice of prior distributions for the model parameters, we used 50% of the Cochrane Database of Systematic Reviews to specify rival prior distributions for δ and τ. The relative predictive performance of the competing models and rival prior distributions was assessed using the remaining 50% of the Cochrane Database. On average, $\mathcal{H}_1^r$—the model that assumes the presence of a treatment effect as well as across-study heterogeneity—outpredicted the other models, but not by a large margin. Within $\mathcal{H}_1^r$, predictive adequacy was relatively constant across the rival prior distributions. We propose specific empirical prior distributions, both for the field in general and for each of 46 specific medical subdisciplines. An example from oral health demonstrates how the proposed prior distributions can be used to conduct a BMA meta-analysis in the open-source software R and JASP. The preregistered analysis plan is available at https://osf.io/zs3df/.

KEYWORDS
Bayes factor, empirical prior distribution, evidence

1 | INTRODUCTION

Following Karl Pearson's first quantitative synthesis of clinical trials in 1904, meta-analysis gradually established itself as an irreplaceable method for statistics in medicine.[1] However, over a century later meta-analysis still presents formidable statistical challenges to medical practitioners, especially when the number of primary studies is low. In this case, the estimation of across-study heterogeneity (i.e., across-study standard deviation) τ is problematic;[2-4] moreover, these problematic τ estimates may subsequently distort the estimates of the overall treatment effect size δ.[2,5] The practical relevance of the small sample challenge is underscored by the fact that the median number of studies in a meta-analysis from the Cochrane Database of Systematic Reviews (CDSR) is only 3, with an interquartile range from 2 to 6.[6]

    One statistical method that has been proposed to address the small sample challenge is *Bayesian estimation*, either with weakly informative prior distributions,[7-9] predictive prior distributions based on pseudo-data,[10] or prior distributions

informed by earlier studies.[11] These Bayesian techniques are well suited to estimate the model parameters when the data are scarce; however, by assigning continuous prior distributions to δ and τ, these estimation techniques implicitly assume that the treatment is effective and the studies are not homogeneous.[*] In order to validate these strong assumptions, we may adopt the framework of *Bayesian testing*. Developed in the second half of the 1930s by Sir Harold Jeffreys,[12,13] the Bayesian testing framework seeks to grade the evidence that the data provide for or against a specific value of interest such as $\delta = 0$ and $\tau = 0$ which corresponds to the null model of no effect and the fixed-effect model, respectively. Jeffreys argued that the testing question logically precedes the estimation question, and that more complex models (e.g., the models used for estimation, where δ and τ are free parameters) ought to be adopted only after the data provide positive evidence in their favor: "Until such evidence is actually produced the simpler hypothesis holds the field; the onus of proof is always on the advocate of the more complicated hypothesis."[14(p. 252)]

In the context of meta-analysis, Jeffreys's statistical philosophy demands that we acknowledge not only the uncertainty in the parameter values given a specific model, but also the uncertainty in the underlying models to which the parameters belong. Both types of uncertainty can be assessed and updated using a procedure known as Bayesian model-averaged (BMA) meta-analysis.[†] The BMA procedure applies different meta-analytic models to the data simultaneously, and draws inferences by taking into account all models, with their impact determined by their predictive performance for the observed data.[19-21]

As in other applications of Bayesian statistics, BMA requires that all parameters are assigned prior distributions. However, in contrast to Bayesian estimation, Bayesian testing does not permit the specification of vague or "uninformative" prior distributions on the parameters of interest. Vague prior distributions assign most prior mass to implausibly large values, resulting in poor predictive performance.[13,21,22] In BMA, the relative impact of the models is determined by their predictive performance, and predictive performance in turn is determined partly by the prior distribution on the model parameters. In objective Bayesian statistics,[23] so-called *default prior distributions* have been proposed; these default distributions meet a list of desiderata[24] and are intended for general use in testing. In contrast to this work, here we seek to construct and compare different prior distributions based on existing medical knowledge.[25-27] Specifically, we propose *empirical prior distributions* for δ and τ as applied to meta-analyses of continuous outcomes in medicine. To this aim, we first used 50% of CDSR to develop candidate prior distributions and then used the remaining 50% of CDSR to evaluate their predictive accuracy and that of the associated models.

Below we first outline the BMA approach to meta-analyses and then present the results of a preregistered analysis procedure to obtain and assess empirical prior distributions for δ and τ for the medical field as a whole. Next we propose empirical prior distributions for the 46 specific medical subdisciplines defined by CDSR. Finally, we demonstrate with a concrete example how our results can be applied in practice using the open-source statistical programs R[28] and JASP.[29]

2 | BMA META-ANALYSIS

The standard Bayesian random-effects meta-analysis assumes that a latent individual study effect θ_i is drawn from a Gaussian group-level distribution with mean treatment effect δ and between-study heterogeneity τ.[30,31] Inference then concerns the posterior distributions for δ and τ. This estimation approach allows researchers to answer important questions such as "given that the treatment effect is nonzero, how large is it?"[‡] and "given that there is between-study heterogeneity, how large is it?" Because the standard model assumes that the effect is nonzero, it cannot address the arguably more fundamental questions that involve a hypothesis test,[32,33] such as "how strong is the evidence in favor of the presence or absence of a treatment effect?" and "how strong is the evidence in favor of between-study heterogeneity (between-study standard deviation) vs homogeneity?"[34(p. 274)] Here we outline a BMA approach that allows for both hypothesis testing and parameter estimation in a single statistical framework.[35]

Our generic meta-analysis setup[19,20,36,37] (for the conceptual basis see Jeffreys)[13(p. 276-277 and p. 296)] consists of the following four qualitatively different candidate hypotheses[§]:

---

[*]Under a continuous prior distribution, the prior probability of any particular value (such as $\delta = 0$, which represent the proposition that the treatment is ineffective) is zero.

[†]A different approach is Bayesian model selection based on posterior (out-of-sample) predictive performance such as DIC/WAIC/LOO.[15,16] However, these approaches are unable to provide compelling support in favor of simple models.[17,18]

[‡]More specific versions of this generic question are "given that the treatment effect is nonzero, is it positive or negative?" and "given that the treatment effect is nonzero, what is the posterior probability that it falls in the interval from a to b?"

[§]The terms "hypothesis" and "model" are used interchangeably.

1. the fixed-effect null hypothesis $\mathcal{H}_0^f : \delta = 0, \tau = 0$;
2. the fixed-effect alternative hypothesis $\mathcal{H}_1^f : \delta \sim g(\cdot), \tau = 0$;
3. the random-effects null hypothesis $\mathcal{H}_0^r : \delta = 0, \tau \sim h(\cdot)$;
4. the random-effects alternative hypothesis $\mathcal{H}_1^r : \delta \sim g(\cdot), \tau \sim h(\cdot)$,

where δ represents the group-level mean treatment effect, τ represents the between-study standard deviation (i.e., the treatment heterogeneity), and $g(\cdot)$ and $h(\cdot)$ represent prior distributions that quantify the uncertainty about δ and τ, respectively. The four prior probabilities of the rival hypotheses are denoted by $p(\mathcal{H}_0^f)$, $p(\mathcal{H}_1^f)$, $p(\mathcal{H}_0^r)$, and $p(\mathcal{H}_1^r)$; these may or may not be set to 1/4, reflecting a position of prior equipoise. The main advantage of this framework is that it does not fully commit to any single model on purely a priori grounds. Although in many situations the random-effects alternative hypothesis $\mathcal{H}_1^r$ is an attractive option, it may be less appropriate when the number of studies is small; in addition, as mentioned above, $\mathcal{H}_1^r$ assumes the effect to be present, whereas assessing the degree to which the data undercut or support this assumption may often be one of the primary inferential goals.

In our framework, after specifying the requisite prior distributions $g(\cdot)$ and $h(\cdot)$, the data drive an update from prior to posterior model probabilities, and pertinent conclusions are then drawn using BMA.[38,39] Specifically, the posterior odds of an effect being present, based on observed data y, is the ratio of the sum of posterior model probabilities for $\mathcal{H}_1^f$ and $\mathcal{H}_1^r$ over the sum of posterior model probabilities for $\mathcal{H}_0^f$ and $\mathcal{H}_0^r$:

$$\text{Posterior odds for treatment effect} = \frac{p(\mathcal{H}_1^f|y) + p(\mathcal{H}_1^r|y)}{p(\mathcal{H}_0^f|y) + p(\mathcal{H}_0^r|y)}.$$

In model-averaging terms, this quantity is referred to as the *posterior inclusion odds*, as it refers to the post-data odds of "including" the effect size parameter δ. As a measure of evidence, one may consider the change, brought about by the data, from prior inclusion odds to posterior inclusion odds. This change is known as the *Bayes factor*:[22,33,40]

$$\underbrace{\text{BF}_{10}}_{\substack{\text{Inclusion Bayes factor} \\ \text{for treatment effect}}} = \underbrace{\frac{p(\mathcal{H}_1^f|y) + p(\mathcal{H}_1^r|y)}{p(\mathcal{H}_0^f|y) + p(\mathcal{H}_0^r|y)}}_{\substack{\text{Posterior inclusion odds} \\ \text{for treatment effect}}} \bigg/ \underbrace{\frac{p(\mathcal{H}_1^f) + p(\mathcal{H}_1^r)}{p(\mathcal{H}_0^f) + p(\mathcal{H}_0^r)}}_{\substack{\text{Prior inclusion odds} \\ \text{for treatment effect}}}. \tag{1}$$

One may similarly assess the posterior odds for the presence of heterogeneity by contrasting $\mathcal{H}_0^r$ and $\mathcal{H}_1^r$ vs $\mathcal{H}_0^f$ and $\mathcal{H}_1^f$:

$$\text{Posterior odds for treatment heterogeneity} = \frac{p(\mathcal{H}_0^r|y) + p(\mathcal{H}_1^r|y)}{p(\mathcal{H}_0^f|y) + p(\mathcal{H}_1^f|y)},$$

or one may quantify evidence by the change from prior to posterior inclusion odds:

$$\underbrace{\text{BF}_{\text{rf}}}_{\substack{\text{Inclusion Bayes factor} \\ \text{for treatment heterogeneity}}} = \underbrace{\frac{p(\mathcal{H}_0^r|y) + p(\mathcal{H}_1^r|y)}{p(\mathcal{H}_0^f|y) + p(\mathcal{H}_1^f|y)}}_{\substack{\text{Posterior inclusion odds} \\ \text{for treatment heterogeneity}}} \bigg/ \underbrace{\frac{p(\mathcal{H}_0^r) + p(\mathcal{H}_1^r)}{p(\mathcal{H}_0^f) + p(\mathcal{H}_1^f)}}_{\substack{\text{Prior inclusion odds} \\ \text{for treatment heterogeneity}}}. \tag{2}$$

An attractive feature of this framework is that it allows a graceful data-driven transition from an emphasis on fixed-effect models to random-effects models; with only few studies available, the fixed-effect models likely outpredict the random-effects models and therefore receive more weight. But as studies accumulate, and it becomes increasingly apparent that the treatment effect is indeed random, the influence of the random-effects models will wax and of the fixed-effect models will wane, until inference is dominated by the random-effects models. In addition, the Bayesian framework allows researchers to monitor the evidence as studies accumulate, without the need or want of corrections for optional stopping.[41] This is particularly relevant as the accumulation of studies is usually not under the control of a central agency, and the stopping rule is ill-defined.[42]

Although theoretically promising, the practical challenge for our BMA meta-analysis is to determine appropriate prior distributions for δ and τ. Prior distributions that are too wide will waste prior mass on highly implausible parameter values, thus incurring a penalty for complexity that could have been circumvented by applying a more reasonably peaked

prior distribution. On the other hand, prior distributions that are too narrow represent a highly risky bet; if the effect is not exactly where the peaked prior distribution guesses it to be, the model will incur a hefty penalty for predicting the data poorly, a penalty that could have been circumvented by reasonably widening the prior distribution. There is no principled way around this dilemma: Bayes' rule dictates that evidence is quantified by predictive success, and *predictions* follow from the *prior* predictive distributions.[43] Thus, when the goal is to quantify evidence, the prior distributions warrant careful consideration.[44-48]

Fortunately, the framework presented here contains only two key parameters, δ and τ; moreover, a large clinical literature is available to help guide the specification of reasonable prior distributions. Our goal in this work is to use meta-analyses from the CDSR to create a series of informed prior distributions for both the effect size parameter δ and between-study variance parameter τ.[8,25-27] We will then assess the predictive adequacy of the various models in conjunction with the prior distributions on a hold-out validation set.

3 | CANDIDATE PRIOR DISTRIBUTIONS

We developed and assessed prior distributions for the δ and τ parameters suitable for BMA of continuous outcomes using data from CDSR.[¶] In the remainder of this work, we adopt the terminology of Higgins et al[49]: individual meta-analyses included in each Cochrane review are referred to as "comparisons" and individual studies included in a comparison are referred to as "studies." All of the results were conducted using Cohen's d standardized mean differences (SMD). The analyses presented in this section were executed in accordance with a preregistration protocol (https://osf.io/zs3df/) unless explicitly mentioned otherwise.

In order to assess the predictive adequacy of the various prior distributions and models, we first randomly partitioned the data of the Cochrane reviews in a training and test set. The training set consisted of 3,092 comparisons with a total number of 23,333 individual studies, and the test set consisted of 3,091 comparisons with a total number of 22,117 individual studies. We used the training set to develop prior distributions for the δ and τ parameters and then assessed predictive accuracy using the test set.

3.1 | Developing prior distributions based on the training set

Left panel of Figure 1 outlines the data processing steps performed on the training set (further details are provided in the preregistration protocol, https://osf.io/zs3df/). First, in order to ensure that the training set yields estimates of τ that form a reliable basis for the construction of a prior distribution, we excluded comparisons with fewer than 10 studies. Second, we excluded comparisons for which at least one individual study was reported by the authors of the review to be "non-estimable" (i.e., the effect size of the original study could not be retrieved). Third, we transformed the reported raw mean differences to the SMD using the `metafor` R package.[50] Fourth, to ensure high consistency of the meta-analytic estimates, we re-estimated all comparisons using a frequentist random-effects meta-analytic model with restricted maximum likelihood estimator using the `metafor` R package.[50] These steps resulted in a final training set featuring 423 comparisons containing a total of 8,044 individual studies. The histograms and tick marks in Figure 2 display the δ and τ estimates from each comparison in the training set.[#]
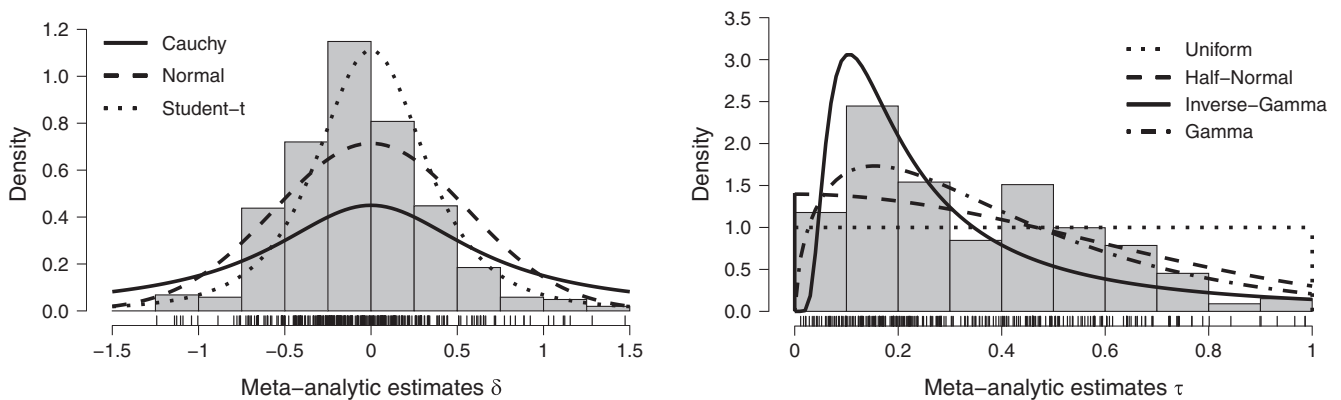
To develop candidate prior distributions for parameters δ and τ, we used the maximum likelihood estimator implemented in the `fitdistrplus` R package[51] to fit several distributions to the frequentist meta-analytic estimates from the training set. For the δ parameter, we considered normal and Student's t distributions fitted to the training set and compared them to an uninformed Cauchy distribution with scale $1/\sqrt{2}$ (a default choice in the field of psychology).[52] For the τ parameter, we considered half-normal, inverse-gamma, and gamma distributions fitted to the training set and compared

---

[¶]We identified systematic reviews in the CDSR through PubMed, limiting the period to January 2000 to May 2020. For that we used the NCBI's EUtils API with the following query: "Cochrane Database Syst Rev"[journal] AND ("2000/01/01"[PDAT]: "2020/05/31"[PDAT]). For each review, we downloaded the XML meta-analysis table file (rm5-format) associated with the review's latest version. We extracted the tables with continuous outcomes (i.e., MD and SMD) from these rm5-files with a custom PHP script. We labeled the tables with the Cochrane Review Group's taxonomy list for subfield analysis.

[#]As specified in the preregistration protocol, we assumed that τ estimates lower than 0.01 are representative of $\mathcal{H}^f : \tau = 0$, and therefore these estimates were not used to determine candidate prior distributions for τ.
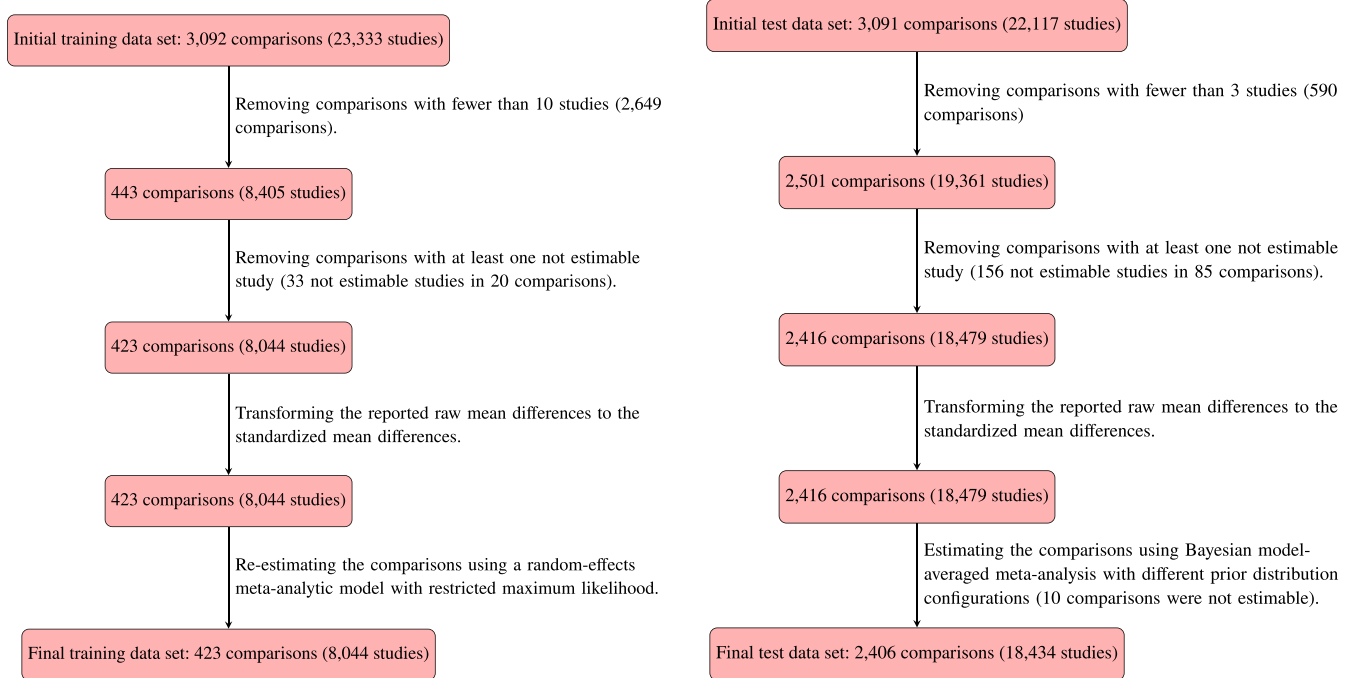
FIGURE 1 Flowchart of the study selection procedure and data processing steps for the training data set (left) and the test data set (right) [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 2 Frequentist effect sizes estimates and candidate prior distributions from the training data set. Histogram and tick marks display the estimated effect size estimates (left) and between-study standard deviation estimates (right), whereas lines represent three associated candidate prior distributions for the population effect size parameter δ (left) and four candidate prior distributions for the population between-study standard deviation τ (right; see Table 1). Twelve effect sizes outside of the ± 1.5 range are not shown and twenty-four τ estimates larger than 1 and sixty-eight τ estimates lower than 0.01 are not shown

them to an uninformed uniform distribution on the range from 0 to 1.[45] The resulting distributions are summarized in Table 1 and their fit to the training set is visualized in Figure 2.

3.2 | Assessing prior distributions based on the test set

Right panel of Figure 1 outlines the data processing steps performed on the test set. Similarly to the training set, we removed non-estimable comparisons and transformed all effect sizes to SMD. However, in contrast to the training set, we retained all comparisons that feature at least 3 studies: there is no reason to limit the assessment of predictive performance

TABLE 1 Candidate prior distributions for the δ and τ parameters as obtained from the training set

Parameter δ	Parameter τ
$\delta \sim \text{Cauchy}(0, 1/\sqrt{2})$	$\tau \sim \mathcal{U}(0, 1)$
$\delta \sim \mathcal{N}(0, 0.56^2)$	$\tau \sim \mathcal{N}_+(0, 0.57^2)$
$\delta \sim \mathcal{T}(0, 0.33, 3)$	$\tau \sim \text{Inv-Gamma}(1.26, 0.24)$
	$\tau \sim \text{Gamma}(1.59, 0.26)$

Note: The inverse-gamma and gamma distributions follow the shape and scale parameterization and the Studen-*t* distributions follow the location, scale, and degrees of freedom parametrization. See Figure 2.

TABLE 2 Ranking totals for each prior configuration based on the 2,406 comparisons in the test set

		Rank											
Prior δ	**Prior τ**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1. $\delta \sim \text{Cauchy}(0, 1/\sqrt{2})$	$\tau \sim \mathcal{U}(0, 1)$	67	74	35	92	51	56	128	89	91	111	128	1484
2. $\delta \sim \text{Cauchy}(0, 1/\sqrt{2})$	$\tau \sim \mathcal{N}_+(0, 0.57^2)$	7	54	39	38	62	82	103	170	134	327	1390	0
3. $\delta \sim \text{Cauchy}(0, 1/\sqrt{2})$	$\tau \sim \text{Inv-Gamma}(1.26, 0.24)$	54	17	46	64	35	55	109	455	706	163	133	569
4. $\delta \sim \text{Cauchy}(0, 1/\sqrt{2})$	$\tau \sim \text{Gamma}(1.59, 0.26)$	9	23	41	51	73	55	62	84	493	1249	261	5
5. $\delta \sim \mathcal{N}(0, 0.56^2)$	$\tau \sim \mathcal{U}(0, 1)$	282	180	50	139	75	101	155	853	443	59	39	30
6. $\delta \sim \mathcal{N}(0, 0.56^2)$	$\tau \sim \mathcal{N}_+(0, 0.57^2)$	8	110	350	217	253	470	933	38	13	9	5	0
7. $\delta \sim \mathcal{N}(0, 0.56^2)$	$\tau \sim \text{Inv-Gamma}(1.26, 0.24)$	247	226	221	724	143	120	106	229	139	191	66	4
8. $\delta \sim \mathcal{N}(0, 0.56^2)$	$\tau \sim \text{Gamma}(1.59, 0.26)$	123	189	128	230	677	808	196	22	15	6	8	4
9. $\delta \sim \mathcal{T}(\mu = 0, \sigma = 0.33, \nu = 3)$	$\tau \sim \mathcal{U}(0, 1)$	227	162	79	283	549	303	304	166	72	55	94	112
10. $\delta \sim \mathcal{T}(\mu = 0, \sigma = 0.33, \nu = 3)$	$\tau \sim \mathcal{N}_+(0, 0.57^2)$	30	197	1051	257	233	199	111	93	99	83	53	0
11. $\delta \sim \mathcal{T}(\mu = 0, \sigma = 0.33, \nu = 3)$	$\tau \sim \text{Inv-Gamma}(1.26, 0.24)$	1122	207	111	137	60	47	101	140	92	52	155	182
12. $\delta \sim \mathcal{T}(\mu = 0, \sigma = 0.33, \nu = 3)$	$\tau \sim \text{Gamma}(1.59, 0.26)$	230	967	265	174	195	110	98	67	109	101	74	16

Note: The numbers indicate how many times a specific prior configuration attained a specific posterior probability rank amongst the 12 possible prior configurations. Rank "1" represents the best performance. Note that these rankings are conditional on assuming the meta-analytic model $\mathcal{H}_1^r$ (i.e., the posterior probabilities of the other meta-analytic models are not considered).

to comparisons with at least 10 studies. These data processing steps resulted in a final test set consisting of 2,416 comparisons containing a total of 18,479 individual studies. The median number of studies in a comparison was 5 with an interquartile range from 3 to 9.

For each possible pair of candidate prior distributions depicted in Table 1,[∥] we computed posterior model probabilities and model-averaged Bayes factors with the `metaBMA R` package,[53] which uses numerical integration and bridge sampling.[54-56]

3.2.1 | Performance of prior distribution configurations under $\mathcal{H}_1^r$

In the first analysis, we evaluate the predictive performance associated with the different prior distribution configurations as implemented in $\mathcal{H}_1^r$, that is the random-effects model that allows both δ and τ to be estimated from the data. Specifically, under $\mathcal{H}_1^r$ there are $3 \times 4 = 12$ prior configurations and each is viewed as a model yielding predictions. The prior probability of each prior configuration is $1/12 \approx 0.083$ and the predictive accuracy of each prior configuration is assessed with the 2,406 comparisons from the test set. Table 2 lists the 12 different prior configurations and summarizes the number of times their posterior probability was ranked $1, 2, \ldots, 12$. The results show that informed configurations

---

[∥]We failed to estimate all BMA meta-analytic models for 10 comparisons due to large outliers.
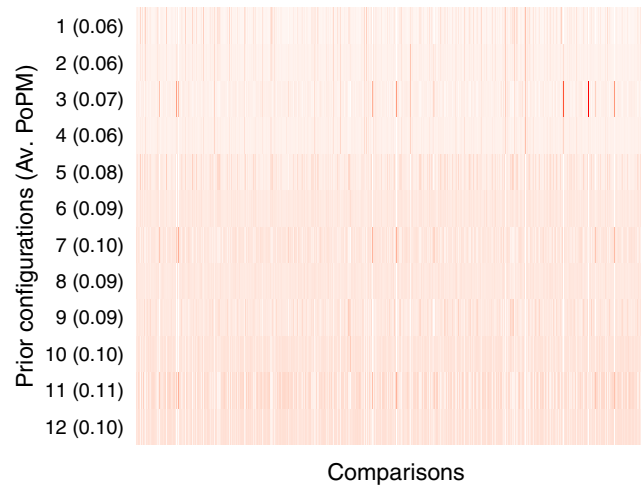
FIGURE 3 Average posterior probabilities (AV. PoMP) for each of the 12 prior configurations under $\mathcal{H}_1^r$ for all 2,406 test-set comparisons individually. For each comparison, the color gradient ranges from white (low posterior probability) to dark red (high posterior probability). The numbers in parentheses are the averaged posterior probabilities across all 2,406 comparisons (conditional on $\mathcal{H}_1^r$). The prior probability for each configuration is $1/12 \approx 0.083$. See also Table 2 [Colour figure can be viewed at wileyonlinelibrary.com]

generally outperformed the uninformed configurations (i.e., the Cauchy$(0, 1/\sqrt{2})$ distribution on δ and the $\mathcal{U}(0, 1)$ distribution on τ). The worst ranking performance was obtained with prior configuration 1 (i.e., uniformed distributions for both δ and τ). Prior configurations 2, 3, 4, 5, and 9 feature an uninformed prior distribution on either δ or τ, and also did not perform well in terms of posterior rankings. The same holds for prior distribution configurations with the half-normal prior distribution for the τ parameter (i.e., prior configurations 2, 6, and 10). The best performing prior distribution configurations (i.e., numbers 7, 11, and 12) used more data-driven prior distributions for both δ (i.e., fitted normal or t-distributions) and τ (i.e., fitted inverse-gamma or gamma).

Figure 3 displays the posterior probability for each of the 12 prior distribution configurations across the 2,406 comparisons. The color gradient ranges from white (representing low posterior probability) to dark red (representing high posterior probability). Figure 3 shows that, on average, the different prior distribution configurations perform similarly. As suggested by the posterior rankings from Table 2, configuration 1 predicted the data relatively poorly, resulting in an average posterior probability of 0.06; in contrast, configuration 11 predicted the data relatively well, resulting in an average posterior probability of 0.11. However, these posterior probabilities differ only modestly from the prior probability of $1/12 \approx 0.083$, and the Bayes factor associated with the comparison of posterior probabilities of 0.11 and 0.06 is less than 2.

In sum, among the 12 prior configurations under $\mathcal{H}_1^r$ the best predictive performance was consistently obtained by data-driven priors, and the worst predictive performance was obtained by uninformed priors (cf. the posterior rankings from Table 2). However, the extent of this predictive advantage is relatively modest: starting from a prior probability of $1/12 \approx 0.083$, the worst prior configuration has an average posterior probability of 0.06, and the best prior configuration has average posterior probability of 0.11 (cf. Figure 3).**

3.2.2 | Posterior probability of the four model types

In the second analysis, we evaluate the predictive performance of the four meta-analytic model types (i.e., $\mathcal{H}_0^f$, $\mathcal{H}_1^f$, $\mathcal{H}_0^r$, and $\mathcal{H}_1^r$) by model-averaging across all prior distribution configurations, separately for each of the 2,406 comparisons. Table 3 shows the prior model probabilities obtained by first assigning probability 1/4 to each of the four model types, and then spreading that probability out evenly across the constituent prior distribution configurations.[33(p. 47)]

For any particular comparison, a model type's model-averaged posterior probability is obtained by summing the posterior probabilities of the constituent prior distribution configurations. For example, the model-averaged posterior

---

**This modest change in plausibility is consistent with the fact that for overlapping prior distributions, there is an asymptotic limit to the evidence that is given by the ratio of prior ordinates evaluated at the maximum likelihood estimate.[57]

TABLE 3 Overview of the prior probability assignment to the different models and prior distribution configurations

Model	Prior model probability	Effect size δ	Heterogeneity τ	Prior configuration probability
$\mathcal{H}_0^f$	1/4	$\delta = 0$	$\tau = 0$	1/4
$\mathcal{H}_1^f$	1/4	$\delta \sim \text{Cauchy}(0, 1/\sqrt{2})$	$\tau = 0$	1/12
		$\delta \sim \mathcal{N}(0, 0.56^2)$	$\tau = 0$	1/12
		$\delta \sim \mathcal{T}(0, 0.33, 3)$	$\tau = 0$	1/12
$\mathcal{H}_0^r$	1/4	$\delta = 0$	$\tau \sim \mathcal{U}(0, 1)$	1/16
		$\delta = 0$	$\tau \sim \mathcal{N}_+(0, 0.57^2)$	1/16
		$\delta = 0$	$\tau \sim \text{Inv-Gamma}(1.26, 0.24)$	1/16
		$\delta = 0$	$\tau \sim \text{Gamma}(1.59, 0.26)$	1/16
$\mathcal{H}_1^r$	1/4	$\delta \sim \text{Cauchy}(0, 1/\sqrt{2})$	$\tau \sim \mathcal{U}(0, 1)$	1/48
		$\delta \sim \text{Cauchy}(0, 1/\sqrt{2})$	$\tau \sim \mathcal{N}_+(0, 0.57^2)$	1/48
		$\delta \sim \text{Cauchy}(0, 1/\sqrt{2})$	$\tau \sim \text{Inv-Gamma}(1.26, 0.24)$	1/48
		$\delta \sim \text{Cauchy}(0, 1/\sqrt{2})$	$\tau \sim \text{Gamma}(1.59, 0.26)$	1/48
		$\delta \sim \mathcal{N}(0, 0.56^2)$	$\tau \sim \mathcal{U}(0, 1)$	1/48
		$\delta \sim \mathcal{N}(0, 0.56^2)$	$\tau \sim \mathcal{N}_+(0, 0.57^2)$	1/48
		$\delta \sim \mathcal{N}(0, 0.56^2)$	$\tau \sim \text{Inv-Gamma}(1.26, 0.24)$	1/48
		$\delta \sim \mathcal{N}(0, 0.56^2)$	$\tau \sim \text{Gamma}(1.59, 0.26)$	1/48
		$\delta \sim \mathcal{T}(0, 0.33, 3)$	$\tau \sim \mathcal{U}(0, 1)$	1/48
		$\delta \sim \mathcal{T}(0, 0.33, 3)$	$\tau \sim \mathcal{N}_+(0, 0.57^2)$	1/48
		$\delta \sim \mathcal{T}(0, 0.33, 3)$	$\tau \sim \text{Inv-Gamma}(1.26, 0.24)$	1/48
		$\delta \sim \mathcal{T}(0, 0.33, 3)$	$\tau \sim \text{Gamma}(1.59, 0.26)$	1/48

probability for $\mathcal{H}_1^r$ is obtained by summing the posterior probabilities for the 12 possible prior configurations, each of them associated with prior probability 1/48 (cf. Table 3).

Table 4 lists the four model types and summarizes the number of times their model-averaged posterior probability was ranked $1, 2, \ldots, 4$. The results show that, across all comparisons, complex models generally received more support than simple models. The model that predicted the data best was $\mathcal{H}_1^r$, the random-effects model that assumes the presence of an effect; the model that predicted the data worst was $\mathcal{H}_0^f$, the fixed-effect model that assumes the absence of an effect. However, even $\mathcal{H}_0^f$ outpredicted the other three model types in $662/2406 \approx 28\%$ of comparisons. Table 4 also shows the model-averaged posterior model probability across all comparisons. In line with the ranking results, the average probability for $\mathcal{H}_0^f$ decreased from 0.25 to 0.19, whereas that for $\mathcal{H}_1^r$ increased from 0.25 to 0.36. Nevertheless, the support for $\mathcal{H}_1^r$ across all comparisons is not overwhelming and does not appear to provide an empirical license to ignore $\mathcal{H}_0^f$ (or any of the other three model types) from the outset.

Left panel of Figure 4 displays the model-averaged posterior probability for each model type across the 2,406 comparisons. It is apparent that the posterior probability is highest for $\mathcal{H}_1^r$. However, for a substantial number of comparisons (i.e., $(662 + 573 + 406)/2406 \approx 68\%$, cf. Table 4) a different model type performs better. For instance—and in contrast to popular belief—the fixed-effect models $\mathcal{H}_1^f$ and $\mathcal{H}_0^f$ together show the best predictive performance in a slight majority of comparisons (i.e., $(662 + 573)/2406 \approx 51\%$).

3.2.3 | Inclusion Bayes factors

In the third analysis, we assess the inclusion Bayes factors for a treatment effect (cf. Equation 1) and for heterogeneity (cf. Equation 2); that is, we model-average across all prior distribution configurations and across two model types, separately

TABLE 4 Ranking totals for each model type based on the 2,406 comparisons in the test set

Model	Rank				PrMP*	AV. PoMP**
	1	**2**	**3**	**4**		
$\mathcal{H}_0^f$	662	177	183	1382	0.25	0.19
$\mathcal{H}_1^f$	573	334	1235	262	0.25	0.22
$\mathcal{H}_0^r$	406	1158	790	52	0.25	0.24
$\mathcal{H}_1^r$	765	737	196	708	0.25	0.36

Note: The numbers indicate how many times a specific model type attained a specific posterior probability rank. Rank "1" represents the best performance. The rankings reflect predictive adequacy that is model-averaged across the possible prior distribution configurations (cf. Table 3).
*Prior model probability.
**Average posterior model probability.
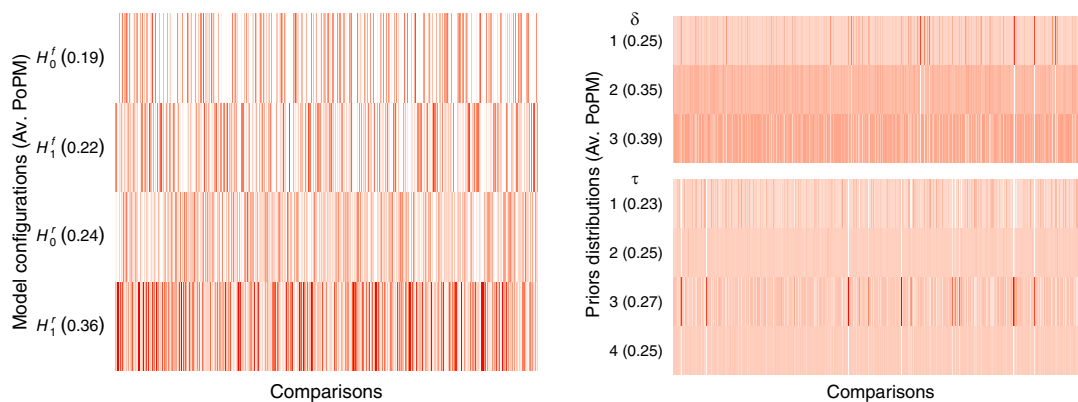
FIGURE 4 Model-averaged posterior probabilities (Av. PoMP) for each of the four model types for all 2,406 test-set comparisons individually (left) and each prior distribution for all 2,406 test-set comparisons individually (right). For each comparison, the color gradient ranges from white (low posterior probability) to dark red (high posterior probability). The numbers in parentheses are the averaged posterior probabilities across all 2,406 comparisons. In the left panel, the prior probability for each model type is 1/4, see also Table 4. In the right panel, the prior probability is 1/3 for each prior distribution on δ, and 1/4 for each prior distribution on τ, see also Table 5 [Colour figure can be viewed at wileyonlinelibrary.com]

for each of the 2,406 comparisons. First, the inclusion BF_{10} for a treatment effect quantifies the evidence that the data provide for the presence vs the absence of a group-level effect, taking into account the model uncertainty associated with whether the effect is fixed or random. The left panel of Figure 5 displays a histogram of the log of the model-averaged BF_{10} for the test set featuring 2,406 comparisons. The histogram is noticeably right-skewed, which affirms the regularity that it is easier to obtain compelling evidence for the presence rather than the absence of an effect.[13(p. 196-197)][††] Evidence for the presence of an effect was obtained in a small majority of the comparisons (i.e., $1336/2406 \approx 55.5\%$).

Second, the inclusion BF_{rf} for heterogeneity quantifies the evidence that the data provide for the presence vs absence of between-study variability, taking into account the model uncertainty associated with whether the group-level effect is present or absent. The right panel of Figure 5 displays a histogram of the log of the model-averaged BF_{rf} for the test set featuring 2,406 comparisons. The right-skew again confirms the regularity: it is easier to find compelling evidence for heterogeneity than for homogeneity. Nevertheless, the data provide evidence for heterogeneity only in a slight majority of $1227/2406 \approx 51.0\%$ of the comparisons.

In sum, the inclusion Bayes factors revealed that in nearly half of the comparisons from the test set, the data provide evidence in favor of the absence of a treatment effect (i.e., 44.5%) and provide evidence in favor of the absence of heterogeneity (i.e., 49.0%). The distribution of the log Bayes factors is asymmetric, indicating that it is easier to obtain compelling

---

[††]This is the case because most prior distributions assign considerable mass to values near the test-relevant one; it is not a universal property of Bayes factors.[58]
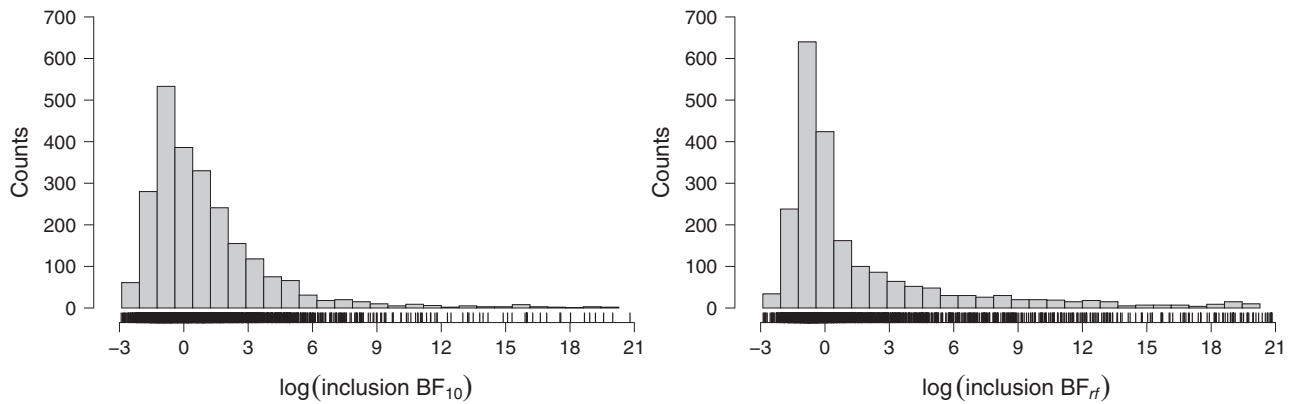
FIGURE 5   Inclusion Bayes factors in favor of the presence of a treatment effect (left) and in favor of the presence of across-study heterogeneity (right) for the 2,406 comparisons in the test set. Not shown are log Bayes factors that exceed 21: twelve log Bayes factors for the presence of a treatment effect and 255 log Bayes factors for the presence of heterogeneity, or are lower than -3: one log Bayes factors for the presence of a treatment effect and four log Bayes factors for the presence of heterogeneity

TABLE 5   Ranking totals for each prior distribution in $\mathcal{H}_1^r$ based on the 2,406 comparisons in the test set

	Rank					
Prior distribution	**1**	**2**	**3**	**4**	**PrMP***	**AV. PoMP****
Parameter δ						
Cauchy$(0, 1/\sqrt{2})$	142	199	2065	–	.33	0.25
$\mathcal{N}(0, 0.56^2)$	655	1727	24	–	.33	0.35
$\mathcal{T}(0, 0.33, 3)$	1609	480	317	–	.33	0.39
Parameter τ						
$\mathcal{U}(0, 1)$	576	83	116	1631	.25	0.23
$\mathcal{N}_+(0, 0.57^2)$	47	772	1587	0	.25	0.25
Inv-Gamma$(1.26, 0.24)$	1418	172	67	749	.25	0.27
Gamma$(1.59, 0.26)$	365	1379	636	26	.25	0.25

Note: The numbers indicate how many times a specific prior distribution attained a specific posterior probability rank. Rank "1" represents the best performance. The rankings reflect predictive adequacy that is model-averaged across the possible prior distribution configurations of the other parameter.
*Prior model probability.
**Average posterior model probability.

evidence for the presence of a treatment effect (rather than for its absence) and for the presence of heterogeneity (rather than for homogeneity).

3.3 | Exploratory analysis: Model-averaging across prior distributions under $\mathcal{H}_1^r$

To further investigate the predictive performance of the prior distributions, we performed one additional analysis that was not preregistered in the original analysis plan. We focused on $\mathcal{H}_1^r$ and evaluated the prior distributions for each parameter by model-averaging across the possible prior distributions for the other parameter. For instance, to obtain the model-averaged posterior probability for the Cauchy prior distribution on the δ parameter (i.e., $\delta \sim \text{Cauchy}(0, 1/\sqrt{2})$), we consider the posterior probability for all 12 possible prior configurations and then sum across the four possible prior distributions on the τ parameter—the four top models listed in the $\mathcal{H}_1^r$ row of Table 3. This way we obtain an assessment of the relative predictive performance of a particular prior distribution, averaging over the uncertainty on the prior distribution for the other parameter.

Table 5 lists the prior distributions and gives the number of times their model-averaged posterior probability attained a particular ranking. Consistent with the results reported earlier, the more data-driven prior distributions generally received

more support than the prior distributions that are less informed. For the δ parameter, the best performing prior distribution was $\delta \sim \mathcal{T}(0, 0.33, 3)$; for the τ parameter, the best performing prior distribution was $\tau \sim$ Inv-Gamma$(1.26, 0.24)$. Although the preference for the data-driven prior distributions is relatively consistent, it is not particularly pronounced, echoing the earlier results. Specifically, Table 5 also shows the model-averaged posterior model probability across all comparisons. For the δ parameter, the t-prior has a model-averaged posterior probability of 0.39 (up from 1/3), but the Cauchy prior retains a non-negligible probability of 0.25. For the τ parameter, the different prior distributions perform even more similarly; on average, the worst prior distribution is $\tau \sim \mathcal{U}(0, 1)$, and yet its model-averaged posterior model probability equals 0.23, down from 1/4 but only a little. Likewise, the on-average best prior distribution is $\tau \sim$ Inv-Gamma$(1.26, 0.24)$, with a model-averaged posterior model probability of 0.27, which is only modestly larger than 1/4.

Right panel of Figure 4 displays the model-averaged posterior probability for each prior distribution across the 2,406 comparisons. The figure confirms that the data-driven prior distributions perform somewhat better than the relatively uninformed prior distributions. The color band is darker red, on average, for the prior distributions with the highest posterior model probabilities, that is, $\delta \sim \mathcal{T}(0, 0.33, 3)$ and $\tau \sim$ Inv-Gamma$(1.26, 0.24)$.

4 | EXPLORATORY ANALYSIS: SUBFIELD-SPECIFIC PRIOR DISTRIBUTIONS

Medical subfields may differ both in the typical size of the effects and in their degree of heterogeneity. In recognition of this fact, we sought to develop empirical prior distributions for δ and τ that are subfield-specific. We differentiated between 47 medical subfields according to the taxonomy of the Cochrane Review Group. Based on their relatively good predictive performance detailed in the previous sections, we selected a t-distribution for the δ parameter (i.e., for subfield i, $\delta_i \sim \mathcal{T}(0, \sigma_i, \nu_i)$) and an inverse-gamma distribution for the τ parameter (i.e., for subfield i, $\tau_i \sim$ Inverse-gamma(α_i, β_i)).

To estimate the parameters of these distributions separately for each subfield, we used the complete data set and proceeded analogously to the training set preparation: we removed comparisons with non-estimable studies, only used comparisons with at least ten studies, re-estimated the comparisons with a restricted maximum likelihood estimator, and removed comparisons with $\tau < 0.01$ estimates. These frequentist estimates were used as input for constructing the data-driven subfield-specific prior distributions. However, since many subfields contain only a limited number of comparisons, we used Bayesian hierarchical estimation with weakly informative priors on the hyperparameters. The hierarchical aspect of the estimation procedure shrinks the estimated parameter values toward the grand mean, a tendency that is more pronounced if the estimated field-specific value is both extreme and based on relatively little information.[59-61] Specifically, we assumed that all field-specific parameters (i.e., σ_i, ν_i, α_i, and β_i) are governed by positive-only normal distributions. For the t-distribution, we assigned positive-only Cauchy$(0, k)$ prior distributions both to the across-field normal mean and to the across-field normal standard deviation, with $k = 1$ for parameter σ and $k = 10$ for parameter ν. For the inverse-gamma distribution, we assigned positive-only Cauchy$(0, 1)$ prior distributions both to the across-field normal mean and to the across-field normal standard deviation for shape parameter α and scale parameter β. The hierarchical models were estimated using the `rstan` R package[62] that interfaces with the Stan probabilistic modeling language.[63] The Stan code is available alongside the supplementary materials at https://osf.io/zs3df/.

Table 6 lists the 46 different subfields (the 47th subfield "Multiple Sclerosis and Rare Diseases of the CNS" featured two comparisons, both of which were excluded based on the $\tau < 0.01$ criterion), the associated number of comparisons and studies, and the estimated distributions for both δ and τ. The scale estimates for the δ parameter show considerable variation, ranging from 0.18 ("Developmental, Psychosocial and Learning Problems") to 0.60 ("Hepato-Biliary").[‡‡] A similar variation is present in the estimated distributions for the τ parameter. Figure 6 visualizes the prior distributions for each subfield.

5 | EXAMPLE: DENTINE HYPERSENSITIVITY

We demonstrate BMA meta-analysis with an example from oral health. Poulsen et al[64] considered the effect of potassium-containing toothpaste on dentine hypersensitivity. Five studies with a tactile outcome assessment were

---

‡‡Notice that the most extreme estimates come from fields with relatively large number of comparisons, as these estimates are less subject to shrinkage.

TABLE 6 Subfield-specific prior distributions for 46 individual topics from the Cochrane Database of Systematic Reviews estimated by hierarchical regression based on the complete data set

Topic	Comparisons	Studies	Prior δ	Prior τ
Acute Respiratory Infections	6	104	$T(0, 0.38, 5)$	Inv-Gamma$(1.73, 0.46)$
Airways	46	815	$T(0, 0.38, 6)$	Inv-Gamma$(2.02, 0.28)$
Anaesthesia	44	661	$T(0, 0.55, 4)$	Inv-Gamma$(1.62, 0.64)$
Back and Neck	13	278	$T(0, 0.37, 5)$	Inv-Gamma$(1.75, 0.57)$
Bone, Joint and Muscle Trauma	32	1221	$T(0, 0.40, 5)$	Inv-Gamma$(1.52, 0.28)$
Colorectal	13	372	$T(0, 0.51, 5)$	Inv-Gamma$(1.64, 0.56)$
Common Mental Disorders	17	264	$T(0, 0.55, 5)$	Inv-Gamma$(1.62, 0.45)$
Consumers and Communication	6	72	$T(0, 0.40, 5)$	Inv-Gamma$(1.56, 0.14)$
Cystic Fibrosis and Genetic Disorders	1	12	$T(0, 0.47, 5)$	Inv-Gamma$(1.70, 0.45)$
Dementia and Cognitive Improvement	9	197	$T(0, 0.45, 5)$	Inv-Gamma$(1.71, 0.44)$
Developmental, Psychosocial and Learning Problems	20	407	$T(0, 0.18, 5)$	Inv-Gamma$(1.43, 0.12)$
Drugs and Alcohol	8	170	$T(0, 0.33, 5)$	Inv-Gamma$(1.89, 0.28)$
Effective Practice and Organisation of Care	10	204	$T(0, 0.39, 5)$	Inv-Gamma$(1.71, 0.35)$
Emergency and Critical Care	9	214	$T(0, 0.39, 5)$	Inv-Gamma$(1.62, 0.29)$
ENT	17	273	$T(0, 0.43, 5)$	Inv-Gamma$(1.85, 0.48)$
Eyes and Vision	14	347	$T(0, 0.40, 6)$	Inv-Gamma$(1.86, 0.41)$
Gynaecological, Neuro-oncology and Orphan Cancer	1	10	$T(0, 0.45, 5)$	Inv-Gamma$(1.67, 0.46)$
Gynaecology and Fertility	14	253	$T(0, 0.38, 5)$	Inv-Gamma$(1.78, 0.46)$
Heart	88	2112	$T(0, 0.42, 5)$	Inv-Gamma$(1.83, 0.47)$
Hepato-Biliary	34	1103	$T(0, 0.60, 4)$	Inv-Gamma$(1.56, 0.58)$
HIV/AIDS	2	23	$T(0, 0.43, 5)$	Inv-Gamma$(1.73, 0.44)$
Hypertension	27	524	$T(0, 0.48, 3)$	Inv-Gamma$(2.01, 0.38)$
Incontinence	17	219	$T(0, 0.33, 6)$	Inv-Gamma$(1.64, 0.36)$
Infectious Diseases	8	150	$T(0, 0.59, 2)$	Inv-Gamma$(1.28, 0.44)$
Inflammatory Bowel Disease	1	12	$T(0, 0.40, 5)$	Inv-Gamma$(1.76, 0.39)$
Injuries	3	54	$T(0, 0.35, 5)$	Inv-Gamma$(1.80, 0.34)$
Kidney and Transplant	39	767	$T(0, 0.54, 5)$	Inv-Gamma$(1.72, 0.53)$
Metabolic and Endocrine Disorders	25	503	$T(0, 0.43, 5)$	Inv-Gamma$(1.71, 0.37)$
Methodology	5	106	$T(0, 0.49, 5)$	Inv-Gamma$(1.72, 0.51)$
Movement Disorders	5	70	$T(0, 0.42, 5)$	Inv-Gamma$(1.88, 0.33)$
Musculoskeletal	32	778	$T(0, 0.45, 6)$	Inv-Gamma$(1.87, 0.38)$
Neonatal	11	259	$T(0, 0.42, 5)$	Inv-Gamma$(1.68, 0.38)$
Oral Health	10	236	$T(0, 0.51, 5)$	Inv-Gamma$(1.79, 0.28)$
Pain, Palliative and Supportive Care	16	283	$T(0, 0.43, 5)$	Inv-Gamma$(1.69, 0.42)$
Pregnancy and Childbirth	32	539	$T(0, 0.33, 5)$	Inv-Gamma$(1.86, 0.32)$
Public Health	2	22	$T(0, 0.33, 5)$	Inv-Gamma$(1.76, 0.23)$
Schizophrenia	21	436	$T(0, 0.29, 4)$	Inv-Gamma$(1.60, 0.27)$
Sexually Transmitted Infections	9	113	$T(0, 0.42, 5)$	Inv-Gamma$(1.70, 0.59)$
Skin	6	85	$T(0, 0.48, 5)$	Inv-Gamma$(1.64, 0.51)$

TABLE 6 (Continued)

Topic	Comparisons	Studies	Prior δ	Prior τ
Stroke	21	357	$\mathcal{T}(0, 0.48, 5)$	Inv-Gamma(1.71, 0.40)
Tobacco Addiction	4	44	$\mathcal{T}(0, 0.44, 4)$	Inv-Gamma(1.73, 0.42)
Upper GI and Pancreatic Diseases	1	12	$\mathcal{T}(0, 0.45, 5)$	Inv-Gamma(1.76, 0.38)
Urology	2	33	$\mathcal{T}(0, 0.44, 5)$	Inv-Gamma(1.73, 0.45)
Vascular	3	35	$\mathcal{T}(0, 0.46, 5)$	Inv-Gamma(1.66, 0.50)
Work	2	24	$\mathcal{T}(0, 0.42, 5)$	Inv-Gamma(1.76, 0.39)
Wounds	7	103	$\mathcal{T}(0, 0.56, 5)$	Inv-Gamma(1.54, 0.41)
Pooled estimate	713	14 876	$\mathcal{T}(0, 0.43, 5)$	Inv-Gamma(1.71, 0.40)

Note: The *t*-distribution follows a location, scale, and degrees of freedom parameterization and the inverse-gamma distribution follows a shape and scale parameterization. See also Figure 6.
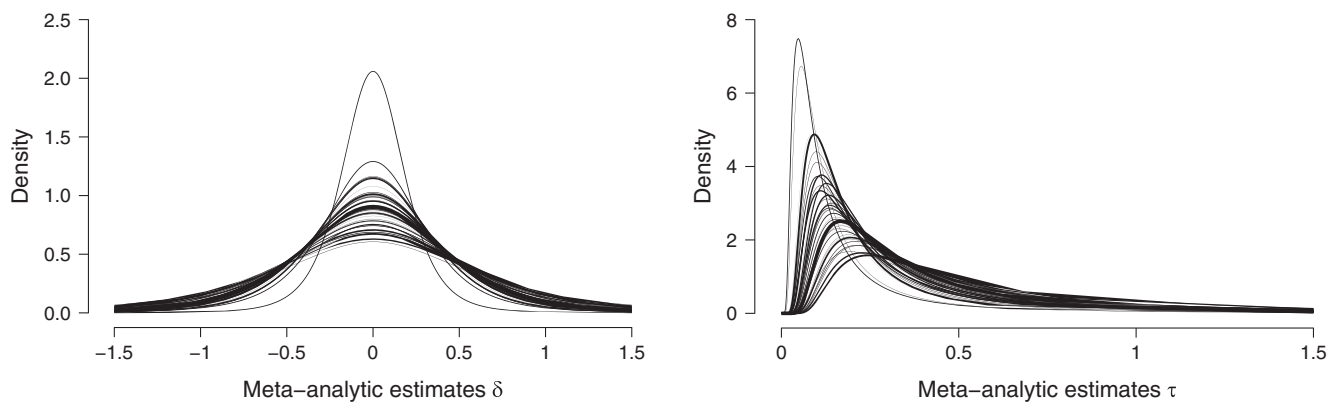

FIGURE 6 Subfield-specific prior distributions for parameter δ (left panel) and parameter τ (right panel) for 46 individual topics from the Cochrane Database of Systematic Reviews estimated by hierarchical regression based on the complete data set. See also Table 6

subjected to a meta-analysis. In their review, Poulsen et al[64] reported a meta-analytic effect size estimate $\delta = 1.19$, 95% CI [0.79, 1.59], $z = 5.86$, $p < 0.00001$ of potassium-containing toothpastes on reducing tactile sensitivity ("Analysis 1.1. Comparison 1 Potassium containing toothpaste (update), Outcome 1 Tactile.").

We reanalyze the Poulsen et al[64] comparison using the BMA meta-analysis implementation in the open-source statistical software package JASP (jasp-stats.org).[29,65-68] The Appendix provides the same analysis in R[28] using the `metaBMA` package.[53] Figure 7 shows the JASP graphical user interface with the left panel specifying the analysis setting and the right panel displaying the default output. After loading the data into JASP, the BMA meta-analysis can be performed by activating the "Meta-Analysis" module after clicking the blue "+" button in the top right corner, choosing "Meta-Analysis" from the ribbon at the top, and then selecting "Bayesian Meta-Analysis" from the drop-down menu. In the left input panel, we move the study effect sizes and standard errors into the appropriate boxes and adjust the prior distributions under the "Prior" tab to match the subfield-specific distributions given in Table 6. Specifically, for the "Oral Health" subfield the prior distributions are $\delta \sim \mathcal{T}(0, 0.51, 5)$ and $\tau \sim$ Inv-Gamma(1.79, 0.28).

The JASP output panel displays the corresponding BMA meta-analysis results. The "Posterior Estimates per Model" table summarizes the estimates and evidence from the fixed-effect models, random-effects models, and finally the model-averaged results. The final row of the table shows an effect size estimate $\delta = 1.082$, 95 % CI [0.686, 1.412] which is slightly lower and more conservative than the one provided by the frequentist random-effects meta-analysis, further quantified with extreme evidence for the presence of an effect, $BF_{10} = 218.53$, and moderate evidence for the presence of heterogeneity, $BF_{rf} = 3.52$. The JASP output panel also presents a forest plot that visualizes the observed effects size estimates from the individual studies, the overall fixed-effect and random-effects meta-analytic estimates and the corresponding model-averaged effect size estimate.
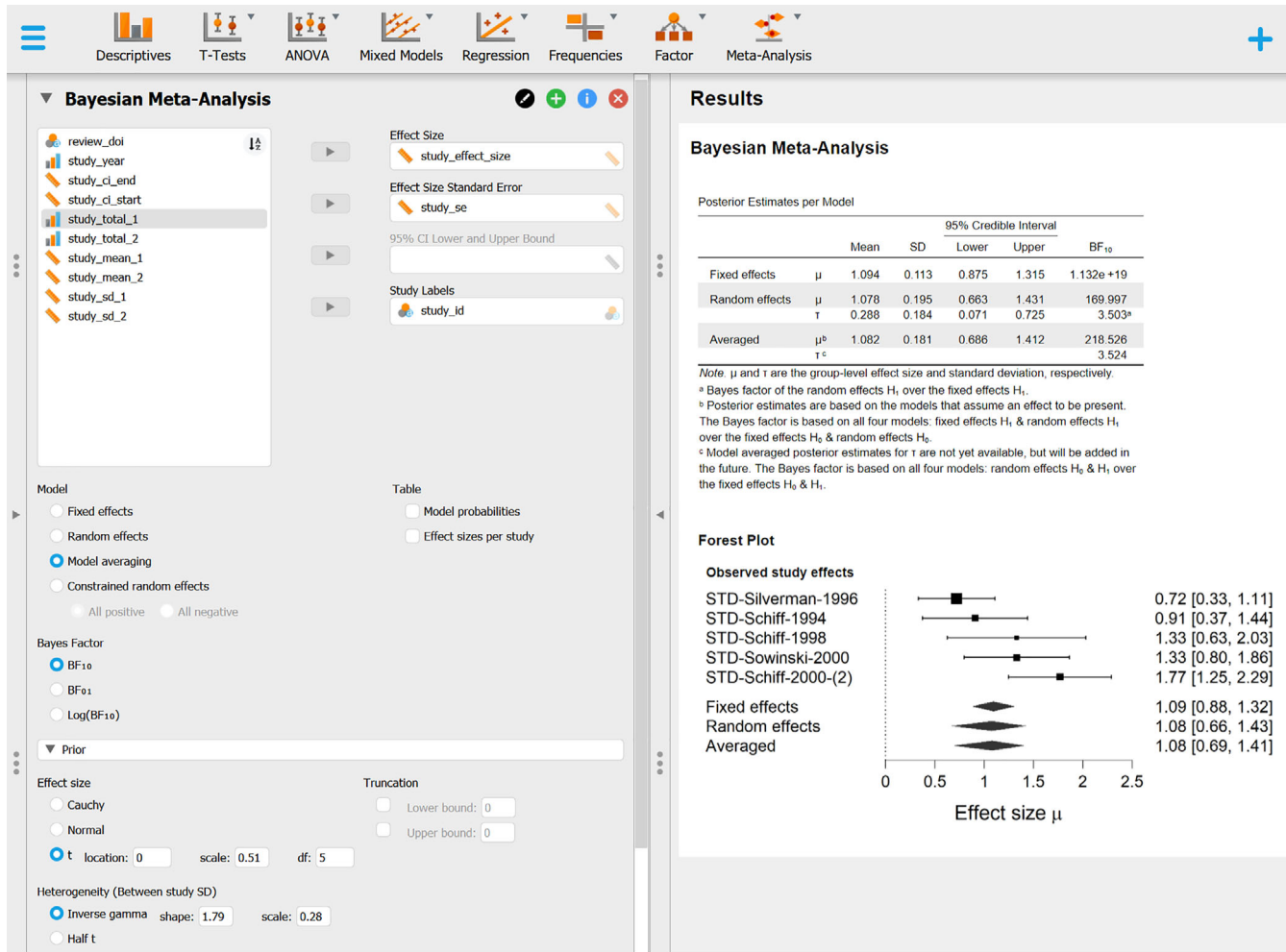
FIGURE 7 JASP screenshot of a Bayesian model-averaged meta-analysis of the Poulsen et al[64] comparison concerning the effect of potassium-containing toothpaste on dentine tactile hypersensitivity. The left input panel shows the specification of the "Oral Health" CDSR subfield-specific prior distributions for effect size δ and heterogeneity τ. The right output panel shows the corresponding results [Colour figure can be viewed at wileyonlinelibrary.com]

The JASP interface provides additional options not discussed here, such as (a) visualizing the prior and posterior distributions; (b) visualizing the estimated effect sizes from individual studies; (c) performing one-sided hypothesis tests; (d) updating evidence sequentially, study-by-study; (e) adding ordinal constraints;[69] and (f) adjustments for publication bias.[70,71]

6 | CONCLUDING COMMENTS

In this article, we introduced BMA meta-analysis for continuous outcomes in medicine. The proposed methodology provides a principled way to integrate, quantify, and update uncertainty regarding both parameters and models. Specifically, the methodology allows researchers to simultaneously test for and estimate effect size and heterogeneity without committing to a particular model in an all-or-none fashion. In BMA meta-analysis, multiple models are considered simultaneously, and inference is proportioned to the support that each model receives from the data. This eliminates the need for stage-wise, multi-step inference procedures that first identify a single preferred model (e.g., a fixed-effect model or a random-effects model) and then interpret the model parameters without acknowledging the uncertainty inherent in the model selection stage. The multi-model approach advocated here also decreases the potential impact of model misspecification.

BMA meta-analysis comes with the usual advantages of Bayesian statistics—the ability to quantify evidence in favor or against any hypothesis (including the null hypothesis), the ability to discriminate between absence of evidence and evidence of absence,[72,73] the ability to monitor evidence as individual studies accumulate,[42] the straightforward

interpretation of the results (i.e., probability statements that refer directly to parameters and hypotheses),[74] and the opportunity to incorporate historical information.[75,76] In this article, our goal was to take advantage of the existing medical knowledge base in order to propose and assess prior distributions that allow for more efficient inference.

Following a preregistered analysis plan, we fitted and assessed different prior distributions for both effect size δ and heterogeneity τ using comparisons of continuous outcomes from the CDSR. We fitted prior distributions based on a training set of randomly selected comparisons, and then evaluated predictive performance based on a test set. The results showed that predictive performance on the test set was relatively similar for the different data-driven prior distributions. Moreover, and in contrast to popular belief and recommendations,[8,9,77,78] we did not find that the random-effects meta-analytic model provided a superior account of the data: the random-effects meta-analytic models outpredicted their fixed-effect counterparts in only 51.0% of comparisons. Although the random-effects alternative hypothesis $\mathcal{H}_1^r$ showed the best predictive performance on average, the data increased its model-averaged posterior probability from 0.25 to only 0.36, leaving 0.64 for the three competing model types (i.e., a model with no heterogeneity, a model without an effect, and a model without both).

Based on the outcome of our preregistered analysis, we used the data from CDSR to develop empirical prior distributions for continuous outcomes in 46 different medical subfields. Finally, we applied BMA meta-analysis with subfield-specific prior distributions to an example from oral health, using the free statistical software packages R and JASP. We believe that the proposed Bayesian methodology provides an alternative perspective on meta-analysis that is informed, efficient, and insightful.

CONFLICT OF INTEREST
František Bartoš, Alexander Ly, and Eric-Jan Wagenmakers declare their involvement in the open-source software package JASP (https://jasp-stats.org), a non-commercial, publicly funded effort to make Bayesian statistics accessible to a broader group of researchers and students.

DATA AVAILABILITY STATEMENT
Data and analysis scripts are publicly available at: https://osf.io/zs3df/files/.

ORCID
František Bartoš https://orcid.org/0000-0002-0018-5573
Quentin F. Gronau https://orcid.org/0000-0001-5510-6943
Bram Timmers https://orcid.org/0000-0002-2863-6997
Willem M. Otte https://orcid.org/0000-0003-1511-6834
Alexander Ly https://orcid.org/0000-0003-3925-3833
Eric-Jan Wagenmakers https://orcid.org/0000-0003-1596-1034

REFERENCES
1. O'Rourke K. An historical perspective on meta-analysis: dealing quantitatively with varying study results. *J R Soc Med*. 2007;100(12):579-582.
2. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Stat Med*. 2001;20(6):825-840.
3. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol*. 2014;14(1):1-12.
4. Gonnermann A, Framke T, Großhennig A, Koch A. No solution yet for combining two independent studies in the presence of heterogeneity. *Stat Med*. 2015;34(16):2476.
5. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Stat Med*. 1995;14(4):395-411.
6. Davey J, Turner RM, Clarke MJ, Higgins JP. Characteristics of meta-analyses and their component studies in the Cochrane database of systematic reviews: a cross-sectional, descriptive analysis. *BMC Med Res Methodol*. 2011;11(1):1-11.
7. Williams DR, Rast P, Bürkner PC. Bayesian meta-analysis with weakly informative prior distributions; 2018.
8. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc A Stat Soc*. 2009;172(1):137-159.
9. Chung Y, Rabe-Hesketh S, Choi IH. Avoiding zero between-study variance estimates in random-effects meta-analysis. *Stat Med*. 2013;32(23):4071-4089.

10. Rhodes KM, Turner RM, White IR, Jackson D, Spiegelhalter DJ, Higgins JP. Implementing informative priors for heterogeneity in meta-analysis using meta-regression and pseudo data. *Stat Med.* 2016;35(29):5495-5511.

11. Higgins JP, Whitehead A. Borrowing strength from external trials in a meta-analysis. *Stat Med.* 1996;15(24):2733-2749.

12. Jeffreys H. Some tests of significance, treated by the theory of probability. *Proc Cambridge Philos Soc.* 1935;31:203-222.

13. Jeffreys H. *Theory of Probability*. 1st ed. Oxford, UK: Oxford University Press; 1939.

14. Jeffreys H. *Scientific Inference*. 1st ed. Cambridge, UK: Cambridge University Press; 1937.

15. Yao Y, Vehtari A, Simpson D, Gelman A. Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Anal.* 2018;13(3):917-1007.

16. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave–one–out cross–validation and WAIC. *Stat Comput.* 2017;27:1413-1432.

17. Gronau QF, Wagenmakers EJ. Limitations of Bayesian leave-one-out cross-validation for model selection. *Comput Brain Behav.* 2019;2:1-11.

18. Gronau QF, Wagenmakers EJ. Rejoinder: more limitations of Bayesian leave-one-out cross-validation. *Comput Brain Behav.* 2019;2:35-47.

19. Gronau QF, van Erp S, Heck DW, Cesario J, Jonas KJ, Wagenmakers EJ. A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: the case of felt power. *Compr Res Soc Psychol.* 2017;2:123-138.

20. Hinne M, Gronau QF, van den Bergh D, Wagenmakers EJ. A conceptual introduction to Bayesian model averaging. *Adv Methods Pract Psychol Sci.* 2020;3:200-215.

21. Gronau QF, Heck DW, Berkhout SW, Haaf JM, Wagenmakers EJ. A primer on Bayesian model-averaged meta-analysis; 2020.

22. Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc.* 1995;90:773-795.

23. Consonni G, Fouskakis D, Liseo B, Ntzoufras I. Prior distributions for objective Bayesian analysis. *Bayesian Anal.* 2018;13:627-679.

24. Bayarri MJ, Berger JO, Forte A, Garcia-Donato G. Criteria for Bayesian model choice with application to variable selection. *Ann Stat.* 2012;40:1550-1577.

25. 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.

26. Pullenayegum BEM. An informed reference prior for between-study heterogeneity in meta-analyses of binary outcomes. *Stat Med.* 2011;30(26):3082-3094.

27. 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.

28. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2004.

29. JASP Team. JASP (Version 0.15) [Computer software]; 2021. https://jasp-stats.org/

30. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials (with discussion). *J R Stat Soc A.* 1994;157:357-416.

31. Sutton AJ, Abrams KR. Bayesian methods in meta–analysis and evidence synthesis. *Stat Methods Med Res.* 2001;10:277-303.

32. Haaf JM, Ly A, Wagenmakers EJ. Retire significance, but still test hypotheses. *Nature.* 2019;567:461.

33. Jeffreys H. *Theory of Probability*. 3rd ed. Oxford, UK: Oxford University Press; 1961.

34. Fisher RA. *Statistical Methods for Research Workers*. 2nd ed. Edinburgh: Oliver and Boyd; 1928.

35. Röver C, Wandel S, Friede T. Model averaging for robust extrapolation in evidence synthesis. *Stat Med.* 2019;38:674-694.

36. Landy JF, Jia M, Ding IL, et al. Crowdsourcing hypothesis tests: making transparent how design choices shape research results. *Psychol Bull.* 2020;146:451-479.

37. Scheibehenne B, Gronau QF, Jamil T, Wagenmakers EJ. Fixed or random? a resolution through model-averaging. reply to Carlsson, Schimmack, Williams, and Burkner. *Psychol Sci.* 2017;28:1698-1701.

38. Jevons WS. *The Principles of Science: A Treatise on Logic and Scientific Method*. London, UK: MacMillan; 1874.

39. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. *Stat Sci.* 1999;14:382-417.

40. Etz A, Wagenmakers EJ. JBS Haldane's contribution to the Bayes factor hypothesis test. *Stat Sci.* 2017;32:313-329.

41. Berger JO, Wolpert RL. *The Likelihood Principle*. 2nd ed. Hayward, CA: Institute of Mathematical Statistics; 1988.

42. Schure TJ, Grünwald P. Accumulation Bias in meta-analysis: the need to consider time in error control. F1000Research; Vol. 8, 2019:962.

43. Evans M. *Measuring Statistical Evidence Using Relative Belief*. Boca Raton, FL: CRC Press; 2015.

44. Gronau QF, Ly A, Wagenmakers EJ. Informed Bayesian *t*-tests. *Am Stat.* 2020;74:137-143.

45. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Stat Med.* 2005;24:2401-2428.

46. O'Hagan A, Buck CE, Daneshkhah A, et al. *Uncertain Judgements: Eliciting Experts' Probabilities*. Chichester, UK: John Wiley & Sons; 2006.

47. O'Hagan A. Expert knowledge elicitation: Subjective but scientific. *Am Stat.* 2019;73:69-81.

48. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JPT. Predicting the extent of heterogeneity in meta–analysis, using empirical data from the Cochrane Database of Systematic Reviews. *Int J Epidemiol.* 2012;41:818-827.

49. Higgins JPT, Thomas J, Chandler J et al. *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, UK: John Wiley & Sons; 2019.

50. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw.* 2010;36(3):1-48.

51. Delignette-Muller ML, Dutang C. fitdistrplus: an R package for fitting distributions. *J Stat Softw.* 2015;64(4):1-34.

52. Morey RD, Rouder JN. BayesFactor: computation of Bayes factors for common designs. R package version 0.9.12-4.2; 2015.

53. Heck DW, Gronau QF, Wagenmakers EJ. metaBMA: Bayesian model averaging for random and fixed effects meta-analysis; 2017. https://cran.r-project.org/package=metaBMA

54. Gronau QF, Sarafoglou A, Matzke D, et al. A tutorial on bridge sampling. *J Math Psychol.* 2017;81:80-97.

55. Gronau QF, Singmann H, Wagenmakers EJ. Bridgesampling: an R package for estimating normalizing constants. *J Stat Softw.* 2020;92.

56. Meng XL, Wong WH. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Stat Sin.* 1996;6:831-860.

57. Ly A, Wagenmakers EJ. Bayes factors for peri-null hypotheses. *Manuscript submitted for publication*; 2021.

58. Johnson VE, Rossell D. On the use of non–local prior densities in Bayesian hypothesis tests. *J R Stat Soc Ser B.* 2010;72:143-170.

59. Lee MD, Wagenmakers EJ. *Bayesian Cognitive Modeling: A Practical Course.* Cambridge, MA: Cambridge University Press; 2013.

60. McElreath R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan.* Boca Raton, FL: Chapman & Hall/CRC Press; 2016.

61. Rouder JN, Lu J. An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychon Bull Rev.* 2005;12:573-604.

62. Stan Development Team. RStan: the R interface to Stan. R package version 2.21.2; 2020. http://mc-stan.org/

63. Stan Development Team. Stan modeling language users guide and reference manual. Version 2.26; 2021. https://mc-stan.org

64. Poulsen S, Errboe M, Mevil YL, Glenny AM. Potassium containing toothpastes for dentine hypersensitivity. *Cochrane Database Syst Rev.* 2006;3.

65. Love J, Selker R, Marsman M, et al. JASP: Graphical statistical software for common statistical designs. *J Stat Softw.* 2019;88:1–17.

66. Ly A, van den Bergh D, Bartoš F, Wagenmakers EJ. Bayesian inference with JASP. *ISBA Bull.* 2021;28:7-15.

67. van Doorn J, van den Bergh D, Böhm U, et al. The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychon Bull Rev.* 2021.

68. Wagenmakers EJ, Love J, Marsman M, et al. Bayesian inference for psychology. Part II: example applications with JASP. *Psychon Bull Rev.* 2018;25:58-76.

69. Haaf JM, Rouder JN. Does every study? Implementing ordinal constraint in meta-analysis; 2020.

70. Maier M, Bartoš F, Wagenmakers EJ. Robust Bayesian meta-analysis: addressing publication bias with model-averaging; 2020.

71. Bartoš F, Maier M, Wagenmakers EJ. Adjusting for publication bias in JASP selection models and robust Bayesian meta-analysis; 2020.

72. Keysers C, Gazzola V, Wagenmakers EJ. Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nat Neurosci.* 2020;23:788-799.

73. Robinson GK. What properties might statistical inferences reasonably be expected to have?—crisis and resolution in statistical inference. *Am Stat.* 2019;73:243-252.

74. O'Hagan A, Forster J. *Kendall's Advanced Theory of Statistics vol. 2B: Bayesian Inference.* 2nd ed. London, UK: Arnold; 2004.

75. Berry DA. Bayesian clinical trials. *Nat Rev Drug Discov.* 2006;5(1):27-36.

76. Hobbs BP, Carlin BP. Practical Bayesian design and analysis for drug and device clinical trials. *J Biopharm Stat.* 2007;18(1):54-80.

77. National Research Council. *Combining Information: Statistical Issues and Opportunities for Research.* Washington, DC: The National Academies Press; 1992.

78. Mosteller F, Colditz GA. Understanding research synthesis (meta-analysis). *Annu Rev Public Health.* 1996;17(1):1-23.

How to cite this article: Bartoš F, Gronau QF, Timmers B, Otte WM, Ly A, Wagenmakers E-J. Bayesian model-averaged meta-analysis in medicine. *Statistics in Medicine.* 2021;1-19. doi: 10.1002/sim.9170

APPENDIX . R CODE FOR THE ORAL HEALTH EXAMPLE

The result for the example analysis can also be obtained with the statistical programming language R.[28] After initializing the R session, the metaBMA R package[53] needs to be installed with the following command (this command needs to be executed only if the package is not already present):

```
install.packages("metaBMA")
```

After the metaBMA has been installed, it must then be loaded into the session, and the data set (containing the effect sizes and standard errors from the individual studies) from the "Tactile Sensitivity.csv" file can be made available, as follows:

```
library("metaBMA")
data <- read.csv("Tactile Sensitivity.csv")
```

In order to perform the BMA meta-analysis with the subfield-specific prior distributions, we first assign the column containing the effect sizes study_effect_size to the y argument, and the column containing standard errors

study_effect_size to the SE argument (the data = data arguments specifies that both of the columns are located in the already loaded data set called data). Then, we specify the prior distributions for the effect size, the d argument, and the heterogeneity, the tau argument, accordingly to the "Oral Health" row in Table 6. Finally, the control argument specifies an additional control argument for the Markov chain Monte Carlo (MCMC) sampler, increasing the target acceptance probability from the default value of 0.80, often required in cases with a small number of studies (see https://mc-stan.org/misc/warnings.html for more information about Stan's warnings and possible solutions):

```
fit_example <- meta_bma(y = study_effect_size, SE = study_se, data = data,
 d = prior("t", c(location = 0, scale = 0.51, nu = 5)),
 tau = prior("invgamma", c(shape = 1.79, scale = 0.28)),
 control = list(adapt_delta = .90))
```

To obtain the numerical summaries of the estimated model, we just execute the name of the object containing the fitted model:

```
fit_example
```

The resulting output corresponds to that given by JASP output (up to MCMC error):

```
> fit_example
### Meta-Analysis with Bayesian Model Averaging ###
 Fixed H0:   d = 0
 Fixed H1:   d ~ 't' (location=0, scale=0.51, nu=5) with support on the interval
             [-Inf,Inf].
 Random H0: d = 0,
             tau ~ 'invgamma' (shape=1.79, scale=0.28) with support on the interval
             [0,Inf].
 Random H1: d ~ 't' (location=0, scale=0.51, nu=5) with support on the interval
             [-Inf,Inf].
             tau ~ 'invgamma' (shape=1.79, scale=0.28) with support on the interval
             [0,Inf].

# Bayes factors:
    (denominator)
(numerator) fixed_H0 fixed_H1 random_H0 random_H1
 fixed_H0   1.00e+00 8.83e-20 4.29e-18 2.52e-20
 fixed_H1   1.13e+19 1.00e+00 4.85e+01 2.85e-01
 random_H0 2.33e+17 2.06e-02 1.00e+00 5.88e-03
 random_H1 3.97e+19 3.50e+00 1.70e+02 1.00e+00

# Bayesian Model Averaging
  Comparison: (fixed_H1 & random_H1) vs. (fixed_H0 & random_H0)
  Inclusion Bayes factor: 218.526
  Inclusion posterior probability: 0.995

# Model posterior probabilities:
         prior    posterior logml
fixed_H0   0.25    1.95e-20   -51.22
fixed_H1   0.25    2.21e-01   -7.34
random_H0 0.25    4.56e-03   -11.23
random_H1 0.25    7.74e-01   -6.09
```

```
# Posterior summary statistics of average effect size:
         mean   sd    2.5%  50%   97.5% hpd95_lower hpd95_upper n_eff    Rhat
averaged 1.085 0.183 0.686 1.091 1.432    0.705        1.446     NA      NA
fixed    1.092 0.118 0.860 1.093 1.325    0.853        1.317    3645.6  1.001
random   1.083 0.200 0.649 1.090 1.451    0.673        1.466    3863.5  1.000
```

The inclusion Bayes factor quantifying the evidence in favor of heterogeneity can be obtained using Equation (2) and output from the "Model posterior probabilities" table.