Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

Bayesian random-effects meta-analysis using the bayesmeta R package

Christian Röver
University Medical Center Göttingen

Abstract

The *random-effects* or *normal-normal hierarchical model* is commonly utilized in a wide range of meta-analysis applications. A Bayesian approach to inference is very attractive in this context, especially when a meta-analysis is based only on few studies. The **bayesmeta** R package provides readily accessible tools to perform Bayesian meta-analyses and generate plots and summaries, without having to worry about computational details. It allows for flexible prior specification and instant access to the resulting posterior distributions, including prediction and shrinkage estimation, and facilitating for example quick sensitivity checks. The present paper introduces the underlying theory and showcases its usage.

Keywords: evidence synthesis, NNHM, between-study heterogeneity.

1. Introduction

1.1. Meta-analysis

Evidence commonly comes in separate bits, and not necessarily from a single experiment. In contemporary science, the careful conduct of systematic reviews of the available evidence from diverse data sources is an effective and ubiquitously practiced means of compiling relevant information. In this context, meta-analyses allow for the formal, mathematical combination of information to merge data from individual investigations to a joint result. Along with qualitative, often informal assessment and evaluation of the present evidence, meta-analytic methods have become a powerful tool to guide objective decision-making (Chalmers *et al.* 2002; Liberati *et al.* 2009; Hedges and Olkin 1985; Hartung *et al.* 2008; Borenstein *et al.* 2009). Applications of meta-analytic methods span such diverse fields as agriculture, astronomy, biology, ecology, education, health research, medicine, psychology, and many more (Chalmers *et al.* 2002).

When empirical data from separate experiments are to be combined, one usually needs to be concerned about the straight one-to-one comparability of the provided results. There may be obvious or concealed sources of *heterogeneity* between different studies, originating e.g. from differences in the selection and treatment of subjects, or in the exact definition of outcomes. Residual heterogeneity may be anticipated in the modeling stage and considered in the estimation process; a common approach is to include an additional variance component to account for between-study variability. On the technical side, the consideration of such a heterogeneity parameter leads to a *random-effects* model rather than a *fixed-effect* model (Hedges and Olkin 1985; Hartung *et al.* 2008; Borenstein *et al.* 2009). Inclusion of a non-zero heterogeneity will generally lead to more conservative results, as opposed to a "naïve" merging of the given data without consideration of potentially heterogeneous data sources.

1.2. The normal-normal hierarchical model (NNHM)

A wide range of problems may be approached using the *normal-normal hierarchical model* (NNHM); this generic random-effects model is applicable when the estimates to be combined are given along with their uncertainties (standard errors) on a real-valued scale. Many problems are commonly solved this way, often after a transformation stage to re-formulate the problem on an appropriate scale. For example, binary data given in terms of contingency tables are routinely expressed in terms of logarithmic odds ratios (and associated standard errors), which are then readily processed via the NNHM.

In the NNHM, measurements and standard errors are modeled via normal distributions, using means and their standard errors as sufficient statistics, while on a second hierarchy level the heterogeneity is modeled as an additive normal variance component as well. The model then has two parameters, the (real-valued) *effect*, and the (positive) *heterogeneity*. If the heterogeneity is zero, then the model reduces to the special case of a *fixed-effect* model (Hedges and Olkin 1985; Hartung *et al.* 2008; Higgins and Green 2011; Borenstein *et al.* 2009). The model and terminology are described in detail in Section 2.1 below. The **bayesmeta** package is based on this simple yet ubiquitous form of the NNHM.

1.3. Analysis within the NNHM framework

The Bayesian solution

The **bayesmeta** package implements a Bayesian approach to inference. Bayesian modeling has previously been advocated and used in the meta-analysis context (Smith *et al.* 1995; Sutton and Abrams 2001; Spiegelhalter 2004; Spiegelhalter *et al.* 2004; Higgins *et al.* 2009; Lunn *et al.* 2013); the difference to the more common "frequentist" methods is that the problem is approached by expressing *states of information* via probability distributions, where the consideration of new data then constitutes an update to a previous information state (Gelman *et al.* 2014; Jaynes 2003; Spiegelhalter *et al.* 1999). A Bayesian analysis allows (and in fact requires) the specification of *prior information*, expressing the *a priori* knowledge, before data are taken into account. Technically this means the definition of a probability distribution, the *prior distribution*, over the unknowns in the statistical model. Once model and prior are specified, the results of a Bayesian analysis are uniquely determined; however, implementing the necessary computations to derive these in practice may still be tricky.

While analysis results will of course depend on the prior setting, the range of reasonable specifications however is usually limited. In the meta-analysis context, non-informative or weakly informative priors for the effect are readily defined, if required. For the between-study heterogeneity an informative specification is often appropriate, especially when only a small number of studies is involved. Interestingly, the number of studies combined in the meta-analyses archived in the *Cochrane Library* is reported by both Davey *et al.* (2011) and Kontopantelis *et al.* (2013) with a median of 3 and a 75% quantile of 6, so that in practice a majority of analyses (including subgroup analyses and secondary outcomes) here is based on as few as 2–3 studies; such cases may not be as unusual as one might expect, at least in medical contexts. Typical meta-analysis sizes may vary across fields; for example, the data collected by van Erp *et al.* (2017) indicate a median number of 12 and first and third quartiles of 5 and 33 studies, respectively, for meta-analyses published in the *Psychological Bulletin*. Standard options for priors are available here, confining the prior probability within reasonable ranges (Spiegelhalter *et al.* 2004). Long-run properties of Bayesian methods have also been compared with common frequentist approaches by Friede *et al.* (2017a,b), with a focus on the common case of very few studies.

Bayesian methods commonly are computationally more demanding than other methods; usually these require the determination of high-dimensional integrals. In some (usually simpler) cases, the necessary integrals can be solved analytically, but it was mostly with the advent of modern computers and especially the development of Markov chain Monte Carlo (MCMC) methods that Bayesian analyses have become more generally tractable (Metropolis and Ulam 1949; Gilks *et al.* 1996). In the present case of random-effects meta-analysis within the NNHM, where only two unkown parameters are to be inferred, computations may be simplified by utilizing numerical integration or importance resampling (Turner *et al.* 2015), both of which require relatively little manual tweaking in order to get them to work. It turns out that computations may be done partly analytically and partly numerically, offering another approach to simplify calculations via the DIRECT algorithm (Röver and Friede 2017). Utilizing this method, the **bayesmeta** package provides direct access to quasi-analytical posterior distributions without having to worry about setup, diagnosis or post-processing of MCMC algorithms. The present paper describes some of the methods along with the usage of the **bayesmeta** package.

Other common approaches

A frequentist approach to inference is largely focused on long-run average operating characteristics of estimators. In this framework, meta-analysis using the NNHM is most commonly done in two stages, where first the heterogeneity parameter is estimated, and then the effect estimate is derived based on the heterogenity estimate. The choice of a heterogeneity estimator poses a problem on its own; a host of different heterogeneity estimators have been described, for a comprehensive summary of the most common ones see e.g. Veroniki *et al.* (2016). A common problem with such estimators of the heterogeneity variance component is that they frequently turn out as zero, effectively resulting in a fixed-effect model, which is usually seen as an undesirable feature. Within this context, Chung *et al.* (2013) proposed a penalized likelihood approach, utilizing a Gamma type "prior" penalty term in order to guarantee non-zero heterogeneity estimates.

The treatment and estimation of heterogeneity in practice has been investigated e.g. by Pullenayegum (2011), Turner *et al.* (2012) and Kontopantelis *et al.* (2013). When looking at large

numbers of meta-analyses published by the Cochrane Collaboration, the majority (57%) of heterogeneity estimates in fact turned out as zero (Turner *et al.* 2012), while the numbers are higher for "small" meta-analyses, and lower for analyses involving many studies (Kontopantelis *et al.* 2013). Meanwhile the choice of analysis method (fixed- or random-effects) also correlates with the number of studies involved, with larger numbers of studies increasing the chances of a random-effects model being employed (Kontopantelis *et al.* 2013). Kontopantelis *et al.* (2013) also compared the fraction of heterogeneity estimates resulting as zero in actual meta-analyses with that obtained from simulation, suggesting that heterogeneity is commonly underestimated or remains undetected.

Once an estimate for the amount of heterogeneity has been arrived at, what is commonly done is to use this as a plug-in estimate and proceed to compute further tests and estimates *conditioning on the heterogeneity estimate* as if its true value were known (Hedges and Olkin 1985; Hartung *et al.* 2008; Borenstein *et al.* 2009). Such a procedure would be warranted if the heterogeneity estimate was estimated with relatively great precision. Notable exceptions here are the methods proposed by Follmann and Proschan (1999); Hartung and Knapp (2001a,b) and Sidik and Jonkman (2002), where the estimation uncertainty in heterogeneity is accounted for (on the technical side resulting in an inflated standard error and a heavier-tailed Student-t distribution to be utilized for deriving tests or confidence intervals), or bootstrap methods (van den Noortgate and Onghena 2005) and parameter estimation in the *generalized inference* framework (Friedrich and Knapp 2013).

Bayesian and frequentist approaches in comparison

While the interpretation of results from a frequentist analysis, especially significance tests and confidence intervals, is commonly challenging and often misunderstood (Morey *et al.* 2016; Hoekstra *et al.* 2014), Bayesian results usually address the actual research question more directly and may be interpreted more intuitively (Jaynes 2003; Spiegelhalter *et al.* 2004; Szucs and Ioannidis 2017; Kruschke and Liddell 2018). On the one hand, frequentist confidence intervals aim to uniformly provide a pre-specified coverage probability *conditionally on any single point in parameter space*, while Bayesian credible intervals account for the prior distribution and consequently provide proper coverage *on average over the prior* (Dawid 1982; see also Appendix A.5, page 39). By their construction, they directly relate to the information on the parameters after considering the data at hand, which is not quite the intention behind classical confidence statements; even if a proper ("frequentist") coverage probability is attained, this may still lead to rather counterintuitive conclusions in the face of actual data (Jaynes 1976; Morey *et al.* 2016).

In some statistical applications, there is little difference between the results from frequentist and Bayesian analyses; often one may be considered a limiting or special case of the other, while interpretations remain somewhat different (Bartholomew 1965; Jaynes 1976; Lindley 1977; Severini 1991; Spiegelhalter *et al.* 1999; Bayarri and Berger 2004). This is not necessarily the case in the present context, as meta-analyses are quite commonly based on few studies, so that certain large-sample asymptotics may not apply. A common misconception, namely that a Bayesian analysis based on a uniform prior generally yielded identical results to a frequentist, purely likelihood-based analysis, is exposed as such here. A crucial feature of meta-analysis problems is that one of the parameters, the heterogeneity, is confined to a bounded parameter space, which sometimes causes problems for frequentist methods (Mandelkern 2002), partly because heterogeneity estimates commonly are not adequately

characterized through a mere point estimate and an associated standard error. The common use of a plug-in estimate for the heterogeneity in frequentist procedures then turns out problematic, as such a strategy usually only makes sense when the estimated parameter is associated with relatively little uncertainty. Choice of a suitable heterogeneity estimator adds to the complication, as, despite their common aim, actual estimates may turn out quite differently, adding some degree of arbitrariness to the inference (Veroniki *et al.* 2016). Within a Bayesian context, these issues do not pose difficulties, and inference on some parameters while accounting for uncertainty in other nuisance parameters is straightforwardly solved through marginalization. This way, uncertainty in heterogeneity is readily accommodated, and since no asymptotic arguments need to be invoked, results are valid also for small sample sizes. For such reasons Bayesian methods have been considered particularly well-suited for hierarchical models in general (Browne and Draper 2006; Kruschke and Liddell 2018), and for meta-analysis problems in particular (Smith *et al.* 1995; Sutton and Abrams 2001). While Bayesian modeling necessitates the specification of a prior probability distribution over all parameters, the range of plausible formulations in a given context is usually limited. Differences in results corresponding to different prior settings are quite natural, as effectively these correspond to differing answers to differently posed questions.

Use of a coherent Bayesian framework also naturally facilitates advanced computations, in which the posterior from a previous analysis constitutes the prior for a subsequent analysis. This is useful for example in sequential meta-analyses (Spence *et al.* 2016), in the design of future experiments (Schmidli *et al.* 2017), or when utilizing historical data in the analysis of clinical trials (Wandel *et al.* 2017).

Implementation

A number of software packages have been developed for frequentist inference within the NNHM framework, for example the *Review Manager (RevMan)*, that is freely available from the Cochrane Collaboration (The Cochrane Collaboration 2014; Higgins and Green 2011), or, within R, the **metafor** and **meta** packages (Viechtbauer 2010; Schwarzer 2007; Schwarzer *et al.* 2015).

Bayesian analyses are usually computationally more demanding, and quite generally these can be approached using MCMC methods (Gilks *et al.* 1996). For example, meta-analysis along with the extension to meta-regression is implemented in the **bmeta** R package (Ding and Baio 2015) by utilizing Gibbs sampling via JAGS (Plummer 2003, 2008). An MCMC approach offers great flexibility, and a number of model variations are also available, for example, a nonparametric generalization of the NNHM in the **bspmma** package (Burr 2012), a generalized approach based on model averaging in the **metaBMA** package (Heck *et al.* 2017), or methods suitable for the special problem of meta-analysis of diagnostic studies in the **bamdit** and **metamisc** packages (Verde 2011; Debray and de Jong 2012).

In certain model constellations, it may be possible to derive exact posterior distributions, as for example implemented in the **mmeta** R package, which utilizes a parametric model for meta-analysis of count data that are provided in terms of contingency tables (Luo *et al.* 2014). Otherwise, inference in a range of model classes may be approached via *integrated nested Laplace approximations (INLA)*, as utilized e.g. in the **meta4diag** package for meta-analysis of diagnostic studies (Guo and Riebler 2015), or in the **nmaINLA** package for network-meta-analysis and -regression (Günhan 2017).

The **bayesmeta** package aims to provide easy access to a fully Bayesian analysis approach within the common NNHM framework. While the use of MCMC methods would be an option here, these usually require a certain amount of expertise and experience in set-up and convergence diagnostics. Also, inference based on MCMC output always contains a certain noise component due to the finite number of samples, which may sometimes constitute a nuisance. Use of the **bayesmeta** packages instantly provides accurate posterior summary figures analogous to output familiar from common (frequentist) meta-analysis output. Posterior distributions may be accessed in quasi-analytical form, and advanced methods, e.g. for prediction or shrinkage estimation, are also provided. Computations are fast and reproducible, allowing for quick sensitivity checks and facilitating larger-scale simulations. Accuracy of the implementation (calibration) may be verified via simulation (see also Appendix A.5).

1.4. Outline

The remaining paper is mostly arranged in two major parts. In the following Section 2, the underlying theory is introduced; first the common NNHM (random-effects) model and its notation are explained, and prior distributions for the two parameters are discussed. Then the resulting likelihood, marginal likelihood and posterior distributions are presented and some general points are introduced.

In Section 3, the actual usage of the **bayesmeta** package is demonstrated; an example data set is introduced, along which the steps of a Bayesian meta-analysis are shown. The determination of summary statistics and plots, as well as possible variations in the analysis setup and the computation of posterior predictive p-values are presented. Section 4 then concludes with a summary.

2. Random-effects meta-analysis

2.1. The normal-normal hierarchical model

The aim is to infer a quantity μ, the *effect*, based on a number k of different measurements which are provided along with their corresponding uncertainties. What is known are the empirical estimates y_i (of μ) that are associated with *known* standard errors σ_i; these constitute the "input data". The ith study's measurement y_i (where $i = 1, \ldots, k$) is assumed to arise as exchangeable and normally distributed around the study's true parameter value θ_i:

$$y_i|\theta_i, \sigma_i \;\sim\; \mathrm{N}(\theta_i, \sigma_i^2), \tag{1}$$

where the variability is due to the sampling error, whose magnitude is given by the (known) standard error σ_i. All studies do not necessarily have identical true values θ_i; in order to accommodate potential between-study heterogeneity in the model, we assume that each study i measures a quantity θ_i that differs from the overall mean μ by another exchangeable, normally distributed offset with variance $\tau^2 \geq 0$:

$$\theta_i|\mu, \tau \;\sim\; \mathrm{N}(\mu, \tau^2). \tag{2}$$

This second model stage implements the *random effects* assumption.

Especially when the study-specific parameters θ_i are not of primary interest, the notation may be simplified by integrating out the "intermediate" θ_i terms and stating the model in its marginal form as

$$y_i|\mu, \tau, \sigma_i \quad \sim \quad \mathrm{N}(\mu, \sigma_i^2 + \tau^2) \tag{3}$$

(Hedges and Olkin 1985; Hartung *et al.* 2008; Borenstein *et al.* 2009, 2010). The two unknowns remaining to be inferred are the mean effect μ and the heterogeneity τ, which is commonly considered a nuisance parameter. The studies' *shrinkage estimates* of θ_i are however sometimes also of interest and may be inferred from the model as well. In the special case of zero heterogeneity ($\tau = 0$), the model simplifies to a fixed-effect model in which the study-specific means θ_i are all identical ($\theta_1 = \ldots = \theta_k = \mu$).

Such two-stage hierarchical models of an overall mean (μ) and study-specific parameters (θ_i) with a random effect for each study are commonly utilized in meta-analysis applications. The simple case of normally distributed error terms at both stages is often convenient and easily tractable, and it also constitutes a good approximation in many cases. So, while the effect here is treated as a continuous parameter, the model is quite commonly utilized to also process different types of data (e.g. logarithmic odds ratios from dichotomous data, etc.) after transformation to a real-valued effect scale (Hedges and Olkin 1985; Hartung *et al.* 2008; Borenstein *et al.* 2009; Viechtbauer 2010; Higgins and Green 2011).

2.2. Prior distributions

General

Among the two unknowns, the effect μ is commonly of primary interest, while the heterogeneity τ usually is considered a nuisance parameter. In order to infer the parameters, we need to specify our prior information about μ and τ in terms of their joint prior probability density function $p(\mu, \tau)$. What exactly constitutes a reasonable prior distribution always depends on the given context (Gelman *et al.* 2014; Spiegelhalter *et al.* 2004; Jaynes 2003). For computational convenience, in the following we assume that we can factor the prior density into independent marginals: $p(\mu, \tau) = p(\mu) \times p(\tau)$. While this may not seem unreasonable, depending on the context, one may also argue in favour of a dependent prior specification (e.g., Senn 2007; Pullenayegum 2011). In the following, we aim to provide a comprehensive overview of popular or sensible options. We will discuss proper as well as improper priors; when using improper priors, the usual care must be taken, as the resulting posterior then may or may not be a proper probability distribution (Gelman *et al.* 2014). The discussed heterogeneity priors are also summarized in Table 1 below.

The effect parameter μ

An obvious choice of a non-informative prior for the effect μ, being a location parameter, is an improper uniform distribution over the real line (Gelman *et al.* 2014; Spiegelhalter *et al.* 2004; Jaynes 2003). A normal prior (with mean μ_p and variance σ_p^2) is a natural choice as an informative prior for the effect μ, and these two are also the cases we will restrict ourselves to for computational convenience and feasibility in the following. The normal prior here constitutes the conditionally conjugate prior distribution for the effect (see also Section 2.5

below). The uninformative uniform prior would also result as the limiting case for increasing prior uncertainty ($\sigma_\mathrm{p} \to \infty$).

A way to guide the choice of a vague prior is by consideration of *unit information priors* (Kass and Wasserman 1995). The idea here is to specify the prior such that its information content (variance) is in some way, possibly somewhat heuristically, equivalent to a single observational unit. For example, if the endpoint is a logarithmic odds ratio (log-OR), a neutral unit information prior may be given by a normal prior with zero mean (centered around an odds ratio of 1, i.e., "no effect") and a standard deviation of $\sigma_\mathrm{p} = 4$. For a derivation, see also Appendix A.1.

The heterogeneity parameter τ: proper, informative priors

Especially since in the meta-analysis context one is commonly dealing with very small numbers of studies k, where not much information on between-study heterogeneity may be expected to be gained from the data, it may be worth while considering the use of informative priors. Depending on the exact context, there often is some information on what values for the heterogeneity are more plausible and which ones are less so, and making use of the present information may make a difference in the end. For example, if the meta-analysis is based on logarithmic odds ratios, it will usually make sense to assume that heterogeneity is unlikely to exceed, say, $\tau = \log(10) \approx 2.3$, which would correspond to roughly an expected factor 10 difference in effects (odds ratios) between trials due to heterogeneity. An extensive discussion of such cases is provided in Spiegelhalter *et al.* (2004, Sec. 5.7). Values for τ between 0.1 and 0.5 here are considered "reasonable", values between 0.5 and 1.0 are "fairly high" and values beyond 1.0 are "fairly extreme". An analogous reasoning would apply for similarly defined outcomes, for example, logarithmic relative risks, logarithmic hazard ratios, or logarithmic variances (Schmidli *et al.* 2017). Consideration of the magnitude of unit information variances (see previous paragraph) may also be helpful in this context, as variability (heterogeneity) between studies will usually be expected to be substantially below the variability between individuals. Along these lines, it is often useful to also consider the implications of prior specifications in terms of the corresponding *prior predictive distributions*; see also Section 3.4 below. The impact of variations of how exactly prior information is implemented in the model may eventually also be checked via sensitivity analyses.

A sensible informative choice for $p(\tau)$ may be the maximum entropy prior for a pre-specified prior expectation $\mathrm{E}[\tau]$, the exponential distribution with rate $\lambda = \frac{1}{\mathrm{E}[\tau]}$ (Jaynes 1968, 2003; Gregory 2005). Log-normal or half-normal prior distributions, e.g. with pre-specified quantiles, may also be useful alternatives. For example, for log-OR (or similar) endpoints, the routine use of half-normal distributions with scale 0.5 or 1.0 has been suggested by Friede *et al.* (2017a,b) and was shown to work well in simulations. In order to gain robustness, one may also consider mixture distributions as informative priors, for example half-Student-t, half-Cauchy, or Lomax distributions, which may be considered heavy-tailed variants of half-normal or exponential distributions (Johnson *et al.* 1994). Use of a heavy-tailed prior distribution will allow for discounting of the prior in favour of the data in case the data appear to be in conflict with prior expectations (O'Hagan and Pericchi 2012; Schmidli *et al.* 2014). The use of weakly informative half-Student-t or half-Cauchy priors may also be motivated via theoretical arguments, as these can be shown to also exhibit favourable frequentist properties (Gelman 2006; Polson and Scott 2012).

Although an inverse-Gamma distribution for an informative prior may seem to be an obvious choice, use of this distribution is generally *not* recommended (Gelman 2006; Polson and Scott 2012). More on informative (as well as uninformative) priors may be found in Spiegelhalter *et al.* (2004), Gelman (2006) and Polson and Scott (2012). Some empirical evidence to consider for informative priors for certain types of endpoints may be found e.g. in Pullenayegum (2011), Turner *et al.* (2012), Kontopantelis *et al.* (2013) and van Erp *et al.* (2017). In particular, Rhodes *et al.* (2015) and Turner *et al.* (2015) derived empirical priors based on data from the *Cochrane database of systematic reviews*; prior information here is expressed in terms of log-normal or log-Student-t distributions.

The heterogeneity parameter τ: proper, 'non-informative' priors

Some "non-informative" proper priors have been proposed that are scale-invariant in the sense that (like the Jeffreys prior discussed below as well) they depend only on the standard errors σ_i. A re-expression of the estimation problem on a different measurement scale would entail a proportional re-scaling of standard errors and so inference effectively remains unaffected. Such priors are discussed e.g. by Spiegelhalter *et al.* (2004, Sec. 5.7.3) and Berger and Deely (1988). Priors like these, however, are somewhat problematic from a logical perspective, as these imply that the prior information on the heterogeneity depended on the accuracy of the individual studies' estimates (Senn 2007).

The following two priors both depend on the harmonic mean s_0^2 of squared standard errors, i.e.,

$$s_0^2 \;=\; \frac{k}{\sum_{i=1}^{k} \sigma_i^{-2}}. \tag{4}$$

The *uniform shrinkage prior* results from considering the "average shrinkage" $S(\tau) = \frac{s_0^2}{s_0^2+\tau^2}$; placing a uniform prior on $S(\tau)$ results in a prior density

$$p(\tau) \;=\; \frac{2\tau s_0^2}{\left(s_0^2+\tau^2\right)^2} \tag{5}$$

for the heterogeneity, which has a median of s_0. For a detailed discussion see e.g. Spiegelhalter *et al.* (2004) or Daniels (1999). A uniform prior in $S(\tau)$ is equivalent to a uniform prior in $1-S(\tau) = \frac{\tau^2}{s_0^2+\tau^2}$ (Spiegelhalter *et al.* 2004), which is an expression very similar to the I^2 measure of heterogeneity due to Higgins and Thompson (2002). Substiting the harmonic mean s_0^2 for their average ($\hat{s}^2$) in the prior density (5) hence yields a uniform prior in I^2.

The *DuMouchel prior* has a similar form and is defined through

$$p(\tau) \;=\; \frac{s_0}{(s_0+\tau)^2}. \tag{6}$$

This implies a log-logistic distribution for the heterogeneity τ that has its mode at $\tau = 0$ and its median at $\tau = s_0$ (Spiegelhalter *et al.* 2004; DuMouchel and Normand 2000).

A *conventional prior* as a proper variation of the Jeffreys prior (see also the closely related variant in (12) below) was given by Berger and Deely (1988) as

$$p(\tau) \;\propto\; \prod_{i=1}^{k} \left(\frac{\tau}{\left(\sigma_i^2+\tau^2\right)^{3/2}} \right)^{1/k}. \tag{7}$$

This prior is in particular intended as a non-informative but proper choice for testing or model selection purposes (Berger and Deely 1988; Berger and Pericchi 2001).

The heterogeneity parameter τ: improper priors

Uninformative priors It is not so obvious what exactly would qualify a prior for τ as "uninformative". One might argue that an uninformative prior should have a probability density function that is monotonically decreasing in τ; another question would be whether the density's intercept $p(\tau = 0)$ should be positive or finite, or what the density's derivative near zero should be. In general, the uninformative prior for a scale parameter in a simple normal model is commonly taken to be uniform in $\log(\tau)$ (and $\log(\tau^2)$) with density $p(\tau) \propto \frac{1}{\tau}$ (Jeffreys 1946; Gelman *et al.* 2014), however, this "log-uniform" prior will not lead to proper, integrable posteriors in the present context (Gelman 2006). Another reasonable choice may be the improper uniform prior on the positive real line, but care must be taken here as usual, as the posterior may end up improper as well; this will not result in a proper posterior when there are only one or two estimates available (i.e., when $k \leq 2$) and an (improper) uniform effect prior is used (Gelman 2006). The uniform prior may be considered a conservative choice in a particular sense, as shown below (Appendix A.2), but on the other hand it may also be considered overly conservative, as it tends to attach a lot of weight to potentially unreasonably large heterogeneity values. Gelman (2006) generally recommends a uniform heterogeneity prior as an uninformative default, *unless* the number of studies k is small, or an informative prior is desired or for other reasons.

One may also argue via certain requirements that an uninformative prior should meet (Jaynes 1968, 2003). For example, it may be reasonable to demand invariance with respect to re-scaling of τ for the prior density $p(\tau)$, leading to a constraint of the form

$$\frac{1}{s}\, p(\tfrac{\tau}{s}) \;=\; f(s)\, p(\tau) \tag{8}$$

for any scaling factor $s > 0$ and some positive-valued function $f(s)$ (i.e., re-scaling should not affect the density's shape). This requirement obviously restricts the range of priors to those with monotonic density functions. It leads to a family of improper prior distributions with densities

$$p(\tau) \;\propto\; \tau^a \tag{9}$$

for $a \in \mathbb{R}$. This family includes (for $a = -1$) the common log-uniform prior for a scale parameter mentioned above, or (for $a = 0$) the uniform prior. But this class also includes further interesting cases, like, for $-1 < a < 0$, a compromise between the above two uniform and log-uniform priors that is (locally) integrable over any interval $[0, u]$ with $0 < u < \infty$ while also being shorter-tailed on the right than the improper uniform prior. An obvious example is (for $a = -0.5$) the prior with monotonically decreasing density function

$$p(\tau) \;\propto\; \frac{1}{\sqrt{\tau}} \tag{10}$$

which corresponds to a uniform prior in $\sqrt{\tau}$. This prior has the unusual property that the prior density, and with that the posterior as well, exhibits a pole (i.e., approaches infinity) at the origin. A value of $a = 1$ would lead to a uniform prior in τ^2, with an even higher preference for large heterogeneity values, which requires at least $k \geq 4$ studies for a proper posterior; this prior is generally not recommended (Gelman 2006).

The Jeffreys prior The non-informative Jeffreys prior (Gelman *et al.* 2014; Jeffreys 1946) for this problem results from the form of the likelihood (see equation (3) or (14) below), or more specifically, the associated expected Fisher information $J(\mu, \tau)$; its probability density function is given by $p(\mu, \tau) \propto \sqrt{\det(J(\mu, \tau))}$. This general form of Jeffreys' prior however is generally not recommended when the set of parameters includes a location parameter as in the present case; see e.g. Jeffreys (1946), Jeffreys (1961, Sec. III.3.10), Berger (1985, Sec. 3.3.3) and Kass and Wasserman (1996, Sec. 2.2). Instead, location parameters are commonly treated as fixed and are conditioned upon (Berger 1985; Kass and Wasserman 1996). In the present case (since μ and τ are orthogonal in the sense that the Fisher information matrix' off-diagonal elements are zero), this leads to Tibshirani's non-informative prior (Tibshirani 1989; Kass and Wasserman 1996, Sec. 3.7), a variation of the general Jeffreys prior, which is of the form

$$p(\tau) \propto \sqrt{\sum_{i=1}^{k} \left(\frac{\tau}{\sigma_i^2 + \tau^2}\right)^2}. \tag{11}$$

In the following, we will consider this variant as the *Jeffreys prior* for the NNHM. This prior also constitutes the *Berger-Bernardo reference prior* for the present problem (Bodnar *et al.* 2016, 2017). The prior is improper, as the right tail asymptotically behaves like $p(\tau) \propto \frac{1}{\tau}$, but it is locally integrable in the left tail with $p(0) = 0$. The resuling posterior is proper as long as $k \geq 2$ (Bodnar *et al.* 2017).

In case of constant standard errors $\sigma_i = \sigma$, the prior's mode is at $\tau = \sigma$. Otherwise the mode tends to be near the smallest σ_i, but the prior may also be multimodal. The Jeffreys prior's dependence on the standard errors σ_i implies that the prior information varies with the precision of the underlying data y_i. With greater precision, lower heterogeneity values are considered plausible. On the other hand, the prior is invariant to the overall scale of the problem (as it scales with the standard errors σ_i) like the proper non-informative priors mentioned above.

Another variation of the Jeffreys prior was given by Berger and Deely (1988) and is defined as

$$p(\tau) \propto \prod_{i=1}^{k} \left(\frac{\tau}{\sigma_i^2 + \tau^2}\right)^{1/k}. \tag{12}$$

This prior is also improper, and it equals the Jeffreys prior in case all standard errors σ_i are identical.

Choice of a prior

The selection of a prior for the effect μ is relatively straightforward. The normal prior's variance allows to vary the width from narrow/informative to wide/uninformative; the improper uniform prior as a limiting case is also available, and this may be the obvious default choice in many cases. Consideration of the unit information prior's width may also help judging the amount of information conveyed by a given informative prior.

The heterogeneity priors discussed above may roughly be categorized in four classes, as shown in Table 1. First of all, one needs to decide whether a proper prior is desired or required. Arguments in favour of a proper prior may include the need for finite marginal likelihoods and Bayes factors in model selection problems, general preference, or a small number (k) of studies.

Table 1: The heterogeneity priors discussed in Section 2.2 may roughly be divided into 4 classes; some of their properties are summarized below.

	proper		improper	
	informative	non-informative	non-informative	scale-invariant
examples	half-normal, half-Student-t, half-Cauchy, log-normal, exponential, ...	uniform shrinkage (5), DuMouchel (6), conventional (7)	Jeffreys (11), Berger-Deely (12)	uniform in τ, uniform in $\sqrt{\tau}$, ... (9)
dependent on σ_i?	no	yes	yes	no
scale-invariant?	no	scales with σ_i	scales with σ_i	yes
k restrictions?	—	$k \geq 1$	$k \geq 2^*$	$k \geq 3^*$

*(less if combined with a proper effect prior)

Among the proper priors one then has the choice between informative distributions, and priors that are supposed to be non-informative, which however depend on the involved studies' standard errors σ_i. The improper priors discussed here are all uninformative in one or another sense; the Jeffreys and Berger-Deely priors also depend on the σ_i, they require at least $k = 2$ available studies, the uniform prior is independent of the σ_i and requires at least 3 studies.

Some prior densities are illustrated in Figure 1. As the choice of a sensible informative prior depends on the context, and some other priors depend on the σ_i values, the priors shown here correspond to the example discussed in Section 3 below. The proper informative half-normal and half-Cauchy priors with scale 0.5 are reasonable choices for log-ORs and similar endpoints. The log-normal prior's parameters are recommended for the type of investigation based on the analysis by Turner *et al.* (2015). The proper uniform shrinkage, DuMouchel and conventional priors depend on the involved studies' standard errors σ_i. The improper Jeffreys and Berger-Deely prior densities do not integrate to a finite value, so their overall scaling is somewhat arbitrary here.

Gelman (2006) generally recommends the improper uniform heterogeneity prior, *unless* the number of studies k is small, or an informative prior is desired or for other reasons. In those cases, an informative prior from the half-Student-t family is recommended, which includes half-Cauchy and half-normal priors as special or limiting cases. Use of the half-Cauchy family is further supported by Polson and Scott (2012) based also on classical frequentist properties. If, for example, the endpoint is a log-OR, then, using the categorization by Spiegelhalter *et al.* (2004, Sec. 5.7), a half-normal prior with scale 0.5 may confine heterogeneity mostly to "reasonable" to "fairly high" values and leave about 5% probability for "fairly extreme" heterogeneity. A larger scale parameter or a heavier-tailed distribution may then serve as a more conservative or more robust reference for a sensitivity check (Friede *et al.* 2017a,b). The Jeffreys prior constitutes another default choice of an uninformative prior; as the *Berger-Bernardo reference prior* it represents the least informative prior in a certain sense (Bodnar *et al.* 2017), and it will yield a proper posterior as long as at least 2 studies are available.

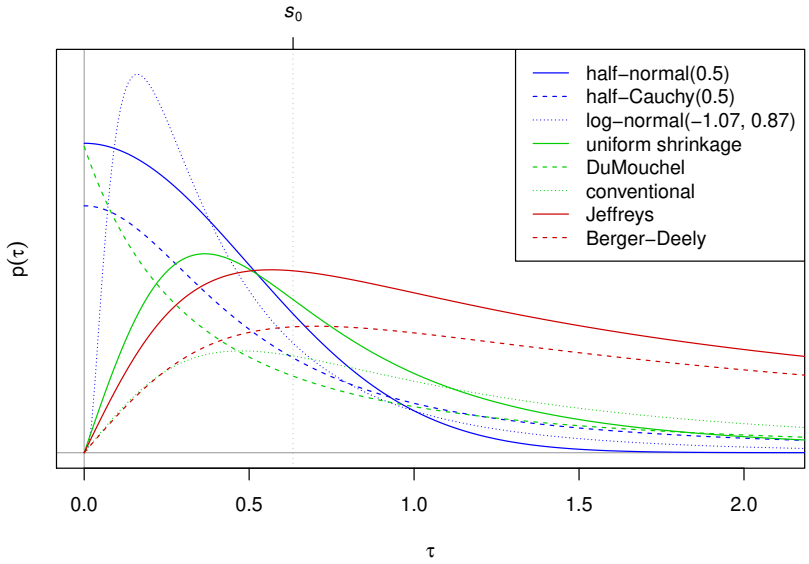

Figure 1: A selection of prior distributions for the example data discussed in Section 3 below. Half-normal and half-Cauchy parameters are reasonable choices for log-OR endpoints. The log-normal parameters are chosen according to Turner *et al.* (2015). The uniform shrinkage and DuMouchel priors are scaled relative to the harmonic mean of squared standard errors s_0^2. The Jeffreys and Berger-Deely priors are improper, so their densities do not integrate to a finite value.

2.3. Likelihood

The form of the likelihood follows from the assumptions introduced in Section 2.1. The NNHM is essentially a simple normal model with unknown mean and an unknown variance component; the resulting likelihood function is given by

$$p(\vec{y}|\mu,\tau,\vec{\sigma}) \;=\; (2\pi)^{-\frac{k}{2}} \times \prod_{i=1}^{k} \frac{1}{\sqrt{\sigma_i^2+\tau^2}} \exp\left(-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma_i^2+\tau^2}\right), \tag{13}$$

where $\vec{y}$ and $\vec{\sigma}$ denote the vectors of k effect measures y_i and their standard errors σ_i. For any practical application it is often more useful to consider the logarithmic likelihood, i.e.,

$$\log(p(\vec{y}|\mu,\tau,\vec{\sigma})) \;=\; -\tfrac{k}{2}\log(2\pi) - \tfrac{1}{2}\sum_{i=1}^{k}\left(\log(\sigma_i^2+\tau^2) + \frac{(y_i-\mu)^2}{\sigma_i^2+\tau^2}\right). \tag{14}$$

2.4. Marginal likelihood

Marginalization

In order to do inference within a Bayesian framework, it is usually necessary to compute integrals involving the posterior distribution (Gelman *et al.* 2014). For example, in a multi-parameter model, one may be interested in the *marginal posterior distribution* or in the

posterior expectation of a certain parameter, both of which result as integrals. Key to the **bayesmeta** implementation is the partly analytical and partly numerical integration over parameter space. In the following, we will derive the marginal posterior distribution of the heterogeneity parameter via the marginal likelihood, and we will later see how marginal and conditional distributions may be utilized to evaluate the required integrals. The likelihood is initially a function of both parameters (μ and τ), and the marginal likelihood of the heterogeneity τ results from integration over the effect μ, using its prior distribution, which we specified to be either uniform or normal.

Uniform prior

Using the improper uniform prior for the effect μ ($p(\mu) \propto 1$), we can derive the marginal likelihood, marginalized over μ,

$$p(\vec{y}|\tau,\vec{\sigma}) \;=\; \int p(\vec{y}|\mu,\tau,\vec{\sigma})\,p(\mu)\,\mathrm{d}\mu. \tag{15}$$

For the NNHM, the integral turns out as

$$p(\vec{y}|\tau,\vec{\sigma}) \;=\; (2\pi)^{-\frac{k-1}{2}} \times \prod_{i=1}^{k} \frac{1}{\sqrt{\sigma_i^2+\tau^2}} \times \exp\!\left(-\frac{1}{2}\frac{\left(y_i-\widehat{\mu}(\tau)\right)^2}{\sigma_i^2+\tau^2}\right) \times \frac{1}{\sqrt{\sum_{i=1}^{k}\frac{1}{\sigma_i^2+\tau^2}}}, \tag{16}$$

where $\widehat{\mu}(\tau)$ is the conditional posterior mean of μ for a given heterogeneity τ. Conditional mean and standard deviation are given by

$$\widehat{\mu}(\tau) \;=\; \mathrm{E}[\mu|\tau,\vec{y},\vec{\sigma}] \;=\; \frac{\sum_{i=1}^{k}\frac{y_i}{\sigma_i^2+\tau^2}}{\sum_{i=1}^{k}\frac{1}{\sigma_i^2+\tau^2}} \qquad \text{and} \qquad \widehat{\sigma}(\tau) \;=\; \sqrt{\mathrm{Var}(\mu|\tau,\vec{y},\vec{\sigma})} \;=\; \sqrt{\frac{1}{\sum_{i=1}^{k}\frac{1}{\sigma_i^2+\tau^2}}}. \tag{17}$$

A derivation is provided in Appendix A.3; the standard deviation $\widehat{\sigma}(\tau)$ will become relevant later on. On the logarithmic scale the marginal likelihood then is:

$$\begin{aligned} &\log(p(\vec{y}|\tau,\vec{\sigma})) \\ =\; &-\frac{1}{2}\left((k-1)\log(2\pi) + \sum_{i=1}^{k}\left(\log(\sigma_i^2+\tau^2) + \frac{\left(y_i-\widehat{\mu}(\tau)\right)^2}{\sigma_i^2+\tau^2}\right) + \log\left(\sum_{i=1}^{k}\frac{1}{\sigma_i^2+\tau^2}\right)\right). \end{aligned} \tag{18}$$

Conjugate normal prior

The normal effect prior here is the *conditionally* conjugate prior distribution, since the resulting conditional posterior (for a given τ value) again is of a normal form. Calculations for the (proper) normal prior for the effect μ work similarly to the previous derivation. Assume the prior for μ is normal with mean μ_p and variance σ_p^2, i.e., it is defined through the probability density function $p(\mu) = \frac{1}{\sqrt{2\pi}\,\sigma_\mathrm{p}}\exp\!\left(-\frac{1}{2}\frac{(\mu-\mu_\mathrm{p})^2}{\sigma_\mathrm{p}^2}\right)$. The necessary integral for the marginal

likelihood then results as

$$p(\vec{y}|\tau,\vec{\sigma}) = \int p(\vec{y}|\mu,\tau,\vec{\sigma})\, p(\mu)\, \mathrm{d}\mu \tag{19}$$

$$= (2\pi)^{-\frac{k+1}{2}} \times \frac{1}{\sqrt{\sigma_\mathrm{p}^2}} \times \prod_{i=1}^{k} \frac{1}{\sqrt{\sigma_i^2+\tau^2}} \times \int \exp\left(-\frac{1}{2}\left[\frac{(\mu-\mu_\mathrm{p})^2}{\sigma_\mathrm{p}^2} + \sum_{i=1}^{k} \frac{(y_i-\mu)^2}{\sigma_i^2+\tau^2}\right]\right) \mathrm{d}\mu. \tag{20}$$

One can see that the prior parameters (μ_p and σ_p) enter in a similar manner as the data points (y_i and σ_i). In analogy to the previous derivation, define the conditional posterior mean and standard deviation

$$\widehat{\mu}(\tau) = \frac{\frac{\mu_\mathrm{p}}{\sigma_\mathrm{p}^2} + \sum_{i=1}^{k} \frac{y_i}{\sigma_i^2+\tau^2}}{\frac{1}{\sigma_\mathrm{p}^2} + \sum_{i=1}^{k} \frac{1}{\sigma_i^2+\tau^2}} \quad \text{and} \quad \widehat{\sigma}(\tau) = \frac{1}{\sqrt{\frac{1}{\sigma_\mathrm{p}^2} + \sum_{i=1}^{k} \frac{1}{\sigma_i^2+\tau^2}}}, \tag{21}$$

and the logarithmic marginal likelihood turns out as

$$\log(p(\vec{y}|\tau,\vec{\sigma})) = -\frac{1}{2}\left(k\log(2\pi) + \log(\sigma_\mathrm{p}^2) + \sum_{i=1}^{k}\log(\sigma_i^2+\tau^2) + \frac{(\mu_\mathrm{p}-\widehat{\mu}(\tau))^2}{\sigma_\mathrm{p}^2} + \sum_{i=1}^{k}\frac{(y_i-\widehat{\mu}(\tau))^2}{\sigma_i^2+\tau^2} + \log\left(\frac{1}{\sigma_\mathrm{p}^2} + \sum_{i=1}^{k}\frac{1}{\sigma_i^2+\tau^2}\right)\right). \tag{22}$$

Note that, comparing equations (18) and (22) (as well as (17) and (21)), as expected, use of the uniform prior constitutes the limiting case of large prior uncertainty ($\sigma_\mathrm{p} \to \infty$).

2.5. Conditional effect posteriors

As long as a uniform or normal prior for the effect μ is used, the effect's conditional posterior distribution for a given heterogeneity value, $p(\mu|\tau,\vec{y},\vec{\sigma})$, again is normal with mean $\widehat{\mu}(\tau)$ and standard deviation $\widehat{\sigma}(\tau)$ as given in equations (17) or (21), respectively (Gelman *et al.* 2014).

Note that the conditional posterior moments (17) are also commonly utilized in frequentist fixed-effect and random-effects meta-analyses. The mean $\widehat{\mu}(\tau)$ constitutes the *conditional maximum likelihood estimate* (of μ), conditional on a particular amount of heterogeneity τ, while $\widehat{\sigma}(\tau)$ gives the corresponding (conditional) standard error. Plugging in $\tau = 0$ yields the *fixed-effect estimate* of μ, while a value $\tau > 0$ yields a *random-effects estimate* (Hedges and Olkin 1985, Sec. 6); see also Section 3.5 below for an example.

2.6. Marginal and joint posterior

Having derived the marginal likelihood $p(\vec{y}|\tau,\vec{\sigma})$ in Section 2.4, the (one-dimensional) marginal posterior density of τ may be computed (up to a normalizing constant) by multiplication with the heterogeneity prior

$$p(\tau|\vec{y},\vec{\sigma}) \propto p(\vec{y}|\tau,\vec{\sigma}) \times p(\tau). \tag{23}$$

This feature was one of the reasons for specifying the priors for μ and τ as independent (see Section 2.2.1). One-dimensional integration can now easily be done numerically for arbitrary priors $p(\tau)$, as long as the resulting posterior is proper.

The effect's conditional posterior $p(\mu|\tau,\vec{y},\vec{\sigma})$ (see Section 2.5) is of particular interest, since the joint posterior may be re-expressed in terms of the conditional as

$$p(\mu,\tau|\vec{y},\vec{\sigma}) \;=\; p(\mu|\tau,\vec{y},\vec{\sigma}) \times p(\tau|\vec{y},\vec{\sigma}). \tag{24}$$

In this formulation, it becomes obvious that the effect's marginal distribution is a continuous *mixture distribution*, in which the normal conditionals $p(\mu|\tau,\vec{y},\vec{\sigma})$ are mixed via the marginal $p(\tau|\vec{y},\vec{\sigma})$ with

$$p(\mu|\vec{y},\vec{\sigma}) \;=\; \int p(\mu,\tau|\vec{y},\vec{\sigma})\,\mathrm{d}\tau \;=\; \int p(\mu|\tau,\vec{y},\vec{\sigma}) \times p(\tau|\vec{y},\vec{\sigma})\,\mathrm{d}\tau \tag{25}$$

(Seidel 2010; Lindsay 1995). This expression allows for easy numerical approximation of posterior integrals of interest. For example, the marginal distribution of the effect μ (the normal mixture) may be approximated by using a discrete grid of τ values and summing up the normal conditionals using weights defined through τ's marginal density:

$$p(\mu) \;=\; \int p(\mu|\tau)\,p(\tau)\,\mathrm{d}\tau \;\approx\; \sum_j p(\mu|\tau_j)\,w_j, \tag{26}$$

where the set of τ_j is appropriately chosen and corresponding "weights" w_j (with $\sum_j w_j = 1$) are based on the marginal $p(\tau)$. With that, it is now relatively straightforward to work with the joint distribution, derive marginals, moments, implement Monte Carlo integration, and so on. A general prescription of how to approach a discrete approximation as sketched in (26) while keeping the accuracy under control is given by the DIRECT algorithm described by Röver and Friede (2017). A few more technical details are also given in Section 2.11 and Appendix A.4 below.

2.7. Predictive distribution

The predictive distribution expresses the posterior knowledge about a "future" observation, i.e., an additional draw θ_{k+1} from the underlying population of studies. This is commonly of interest in order to judge the amount of heterogeneity relative to the estimation uncertainty (Riley *et al.* 2011; Guddat *et al.* 2012; Bender *et al.* 2014), or for extrapolation in the design and analysis of future studies (Schmidli *et al.* 2014). Technically, the predictive distribution $p(\theta_{k+1}|\vec{y},\vec{\sigma})$ is similar to the marginal distribution of the effect μ (see previous section). Conditionally on a given heterogeneity τ, and for the uniform or normal effect prior, the predictive distribution again is normal with moments

$$\mathrm{E}[\theta_{k+1}\,|\,\tau,\vec{y},\vec{\sigma}] \;=\; \widehat{\mu}(\tau) \qquad \text{and} \qquad \mathrm{Var}(\theta_{k+1}\,|\,\tau,\vec{y},\vec{\sigma}) \;=\; \widehat{\sigma}^2(\tau) + \tau^2. \tag{27}$$

2.8. Shrinkage estimates of study-specific means

Sometimes it is of interest to also infer the posterior distributions of the study-specific parameters θ_j. These may e.g. be in the focus if a meta-analysis is performed in order support

the analysis of a particular study by borrowing strength from a number of related studies (Gelman *et al.* 2014; Schmidli *et al.* 2014; Wandel *et al.* 2017). Conditionally on a particular heterogeneity value τ, these distributions are again normal with moments given by

$$\mathrm{E}[\theta_j \,|\, \tau, \vec{y}, \vec{\sigma}] = \frac{\frac{1}{\sigma_j^2} y_j + \frac{1}{\tau^2} \widehat{\mu}(\tau)}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \tag{28}$$

$$\mathrm{Var}(\theta_j \,|\, \tau, \vec{y}, \vec{\sigma}) = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} + \left(\frac{\frac{1}{\tau^2}}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \widehat{\sigma} \right)^2 \tag{29}$$

(Gelman *et al.* 2014, Sec. 5.5). These expressions illustrate the *shrinkage* of posterior estimates towards the common mean as a function of the heterogeneity. Analogously to the effect's posterior and predictive distribution, these conditional moments again allow to approximate each individual θ_i's marginal posterior distribution via a discrete mixture to marginalize over the heterogeneity.

2.9. Credible intervals

Credible intervals derived from a posterior probability distribution may be computed e.g. using the distribution's $\frac{\alpha}{2}$ and $(1-\frac{\alpha}{2})$ quantiles. However, such a simple *central* interval may not necessarily be the most sensible summary of a posterior distribution, especially if it is skewed or extends to the boundary of its parameter space. In such cases, it usually makes more sense to consider the *highest posterior density (HPD) region*, i.e., a $(1-\alpha)$ credible region enclosing the $(1-\alpha)$ posterior probability where the posterior density is largest (Gelman *et al.* 2014). Such a region may be disjoint and hard to determine, but closely related (and identical for unimodal distributions) is the *shortest credible interval*. Both types of intervals, central and shortest, will be considered in the following.

2.10. Posterior predictive checks and p-values

Posterior predictive model checks allow to investigate the fit of a model to a given data set (Gelman *et al.* 1996; Gelman 2003; Gelman *et al.* 2014). The consistency of data and model is explored by comparing the actual data to data sets *predicted* via the posterior distribution. The comparison is usually done graphically, or via suitable summary statistics of actual and predicted data; a discrepancy then is an indicator of a poor model fit.

If the summary statistic is one-dimensional, then the comparison may be formalized by focusing on the fractions of predicted values above or below the actually observed value. This leads to the concept of *posterior predictive p-values*, which are closely related to classical p-values (Meng 1994; Berkhof *et al.* 2000; Gelman 2013; Wasserstein 2016). Posterior predictive p-values have been applied and advocated in a range of contexts, including e.g. educational testing (Sinharay *et al.* 2006), metrology (Kacker *et al.* 2008), psychology (van de Schoot *et al.* 2014) and biology (Chambert *et al.* 2014).

In the context of the NNHM, posterior predictive checks are useful, as they allow to investigate certain hypotheses of interest, like for example $\mu \geq 0$, $\tau = 0$ or $\theta_i = 0$. The posterior predictive distribution *conditional on a particular hypothesis* may then be explored in order to derive a corresponding posterior predictive p-value. The choice of a suitable summary statistic however

may still pose a challenge. The posterior predictive checks here are implemented via Monte Carlo sampling, therefore parts of these procedures are computationally expensive.

2.11. How the `bayesmeta()` function works internally

The `bayesmeta()` function utilizes the fact that in the context of the NNHM the resulting posterior is only 2-dimensional (for now ignoring the θ_i parameters) and may be expressed as a mixture distribution (see (24)) where the heterogeneity's marginal $p(\tau|\vec{y},\vec{\sigma})$ is known, and the effect's conditionals $p(\mu|\tau,\vec{y},\vec{\sigma})$ are all of a normal form. This setup allows to approximate the effect marginal by a discrete mixture (see (26)) while keeping the accuracy under control; the accuracy requirements are formulated via the DIRECT algorithm's two tuning parameters (δ and ϵ) (Röver and Friede 2017).

An example of joint and marginal posterior densities of the two parameters is illustrated in Figure 3 below (see page 23). The joint posterior density (top right) is easily evaluated based on likelihood and prior density, both of which are available in analytical form (see Sections 2.2 and 2.3). The heterogeneity's marginal density (bottom right) is also easily computed, based on marginal likelihood and prior (see (23)); only its normalizing constant needs to be computed numerically (using the `integrate()` function available in R). The CDF is also computed using numerical integration, and the quantile function is evaluated using again the CDF and inverting it via R's `uniroot()` root-finding function.

Now the effect's marginal density (bottom left panel of Figure 3) is approximated by a mixture of a finite number of normal distributions. In terms of equation (26), what is required is a finite set of support points τ_j, the parameters (means and standard deviations) of the associated normal conditionals $p(\mu|\tau_j)$, and the corresponding weights w_j. These are all determined using the DIRECT algorithm, and in the `bayesmeta()` output (see the following section) one can find these in the "`...$support`" element. In the example shown in Figure 3, the effect marginal is based on a 17-component normal mixture; this number of components is sufficient to bound the discrepancy between actual marginal and mixture approximation to amount to a Kullback-Leibler divergence below $\delta = 1\%$. The desired accuracy can be pre-specified via the "`delta`" and "`epsilon`" arguments (Röver and Friede 2017).

Computations related to such discrete, finite mixtures are relatively straightforward; density and CDF are linear combination of the components' (normal) densities and CDFs, random number generation is simple, and moments are also easily derived (Seidel 2010; Lindsay 1995). A few more details on the implementation are given in Appendix A.4. Many of the internal computations heavily rely on numerical integration, root-finding and optimization via R's `integrate()`, `uniroot()`, `optimize()` and `optim()` functions. Accuracy of the eventual implementation is confirmed using simulations in Appendix A.5.

3. Using the bayesmeta package

3.1. General

Before proceeding to an exemplary analysis, we will first introduce an example data set and go through the common procedure of effect size derivation step-by-step. This will serve to introduce some context and generate a set of estimates (y_i) and associated standard errors (σ_i); the subsequent section will then pick up the analysis from that starting point.

3.2. Example data: a systematic review in immunosuppression

Interleukin-2 receptor antagonists (IL-2RA) are commonly used as part of immunosuppressive therapy after organ transplantation. Treatment strategies and responses are different for adults and children, and it was of interest to investigate the effectiveness of IL-2RA in preventing acute rejection (AR) events following liver transplantation in paediatric patients. A systematic literature review was performed, and six controlled studies were found reporting on the occurrence of AR events in paediatric liver transplant recipients (Crins *et al.* 2014).

The binary data on AR events from each of the six studies may be summarized in a 2×2-table as shown in Table 2. The data shown here come from the earliest of the studies found in the review (Heffron *et al.* 2003). Here one can already see that the treatment appears to be effective, as roughly only a quarter of patients in the IL-2RA group experienced an AR event, compared to three quarters in the control group.

In order to compare the effect magnitude between different studies, a common effect measure is computed from each contingency table (for each study i). One such measure is the logarithmic odds ratio (log-OR), comparing the odds of an event in treatment- and control-groups. The log-OR estimate is given by $y_i = \log\left(\frac{a/b}{c/d}\right)$, where a to d are the event counts as defined in Table 2; the corresponding standard error is $\sigma_i = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$. In the above example, the odds ratio is $\frac{14/47}{15/5} = \frac{14}{141} \approx 0.10$; we have $y_1 = \log\left(\frac{14/47}{15/5}\right) = -2.31$ and $\sigma_1 = \sqrt{\frac{1}{14}+\frac{1}{47}+\frac{1}{15}+\frac{1}{5}} = 0.60$. A wide range of other measures is available for contingency tables as well as other types of study outcomes; for example, in the present case one might alternatively be interested in (logarithmic) *relative risks (RR)* ($\log\left(\frac{a/(a+b)}{c/(c+d)}\right)$) instead of the log-ORs (Hedges and Olkin 1985; Hartung *et al.* 2008; Borenstein *et al.* 2009; Viechtbauer 2010; Higgins and Green 2011; Deeks 2002). The original data and derived log-ORs for all

Table 2: The general setup of a 2×2 contingency table for dichotomous outcomes (left) and a concrete example from the paediatric liver transplantation data set (right). Note that one of the three data columns is redundant here, as it may be derived from the remaining two.

	event		
	yes	no	total
treatment	a	b	$n_1 = a+b$
control	c	d	$n_2 = c+d$

	AR event		
	yes	no	total
IL-2RA patients	14	47	61
control patients	15	5	20

Table 3: Data from the immunosuppression example. Each row here summarizes a 2×2 contingency table, the last two columns show the correponding derived log-ORs (y_i) and their associated standard errors (σ_i).

study			IL-2RA group		control group		log-OR	
i	author	year	events (a_i)	total ($n_{1;i}$)	events (c_i)	total ($n_{2;i}$)	y_i	σ_i
1	Heffron *et al.*	2003	14	61	15	20	−2.31	0.60
2	Gibelli *et al.*	2004	16	28	19	28	−0.46	0.56
3	Schuller *et al.*	2005	3	18	8	12	−2.30	0.88
4	Ganschow *et al.*	2005	9	54	29	54	−1.76	0.46
5	Spada *et al.*	2006	4	36	11	36	−1.26	0.64
6	Gras *et al.*	2008	0	50	3	34	−2.42	1.53

six studies from the systematic review are shown in Table 3.

The transplantation data set is also contained in the **bayesmeta** package; the data need to be loaded via the `data()` function:

```
R> require("bayesmeta")
R> data("CrinsEtAl2014")
R> CrinsEtAl2014
```

Effect sizes and standard errors can be calculated from the plain count data either by implementing the corresponding formulas (see above), or, much easier and recommended, by using e.g. the **metafor** package's `escalc()` function:

```
R> require("metafor")
R> crins.es <- escalc(measure="OR",
+    ai=exp.AR.events, n1i=exp.total, ci=cont.AR.events, n2i=cont.total,
+    slab=publication, data=CrinsEtAl2014)
R> crins.es
```

One can see that the `escalc()` function uses a terminology analogous to that in Table 2 to interface with binary outcome data; the "`ai`" input argument corresponds to the a_i table entries (number a of events in the treatment group for each study i), and so on. The output of the `escalc()` function (here: the data frame named "`crins.es`") will then be the original data along with two additional columns named "`yi`" and "`vi`" containing the calculated effect sizes (y_i) and the *squared* (!) standard errors (σ_i^2), respectively.

Note that for computing the 6th study's log-OR (see Table 3), a *continuity correction* was necessary, because one of the contingency table entries was zero (Sweeting *et al.* 2004). For more details on effect size calculation and default behaviour, see also Viechtbauer (2010) or the `escalc()` function's online documentation.

3.3. Performing a Bayesian random-effects meta-analysis

The `bayesmeta()` *function*

In order to perform a random-effects meta-analysis, we need to specify the data, as well as

the prior for the unknown parameters μ and τ (see Section 2.2). For the effect μ we are restricted to normal or uniform priors; here we use a vague prior centered at $\mu_{\mathrm{p}} = 0$, which corresponds to an OR of 1, i.e., no effect. The prior standard deviation we set to $\sigma_{\mathrm{p}} = 4$, corresponding to the vague *unit information prior* (see Section 2.2.2). For the heterogeneity, we use a half-normal prior with scale 0.5, confining the a priori expected heterogeneity to $\tau \leq 0.98$ with 95% probability (i.e., allowing for "fairly extreme" values with only about 5% prior probability).

With the log-ORs computed as in the previous section, we can now execute the analysis using the following call

```
R> ma01 <- bayesmeta(y = crins.es[,"yi"], sigma = sqrt(crins.es[,"vi"]),
+    labels = crins.es[,"publication"], mu.prior.mean = 0, mu.prior.sd = 4,
+    tau.prior = function(t){dhalfnormal(t,scale=0.5)})
```

The first three arguments pass the data (vectors of estimates y_i and standard errors σ_i) and (optionally) a vector of corresponding study labels to the `bayesmeta()` function. Note that the **metafor** package's `escalc()` function returned variances (i.e., *squared* standard errors), while the `bayesmeta()` function's "`sigma`" argument requires the standard errors (i.e., the square root of the variances); hence the additional square-root-transformation here. The following arguments specify the prior mean and standard deviation of the (normal) prior for the effect μ. Finally, the last argument specifies the prior for the heterogeneity τ. While for the effect prior we are restricted to using normal or improper uniform priors, the heterogeneity prior can be of essentially any type. Specification of the heterogeneity prior works via specification of its *prior density function*. While this type of argument specification is somewhat unusual, it is reasonably straightforward, as one can see above. The `dhalfnormal()` function here is the half-normal distribution's density function; see also the corresponding online help (e.g. via entering "`?dhalfnormal`" in R).

Retrieving and processing the "`yi`" and "`vi`" elements (as well as study labels, if available) from an `escalc()` result in general is not complicated, and the `bayesmeta()` function can also do this automatically for any `escalc()` output, including the many types of effect sizes that are available (Viechtbauer 2010). Using simply the `escalc()` function's output as an input, the identical result can be achieved by calling

```
R> ma01 <- bayesmeta(crins.es, mu.prior.mean = 0, mu.prior.sd = 4,
+    tau.prior = function(t){dhalfnormal(t,scale=0.5)})
```

The `bayesmeta()` computations may take up to a few seconds, but with that the main calculations are done, and the essential results are stored in the generated object of class "`bayesmeta`" (here named "`ma01`"). One can inspect the results by printing the returned object:

```
R> ma01

 'bayesmeta' object.

6 estimates:
Heffron (2003), Gibelli (2004), Schuller (2005), Ganschow (2005),
```

```
Spada (2006), Gras (2008)

tau prior (proper):
function(t){dhalfnormal(t,scale=0.5)}

mu prior (proper):
normal(mean=0, sd=4)

ML and MAP estimates:
                     tau        mu
ML joint     0.32581341 -1.578262
ML marginal  0.46441292 -1.578003
MAP joint    0.08690907 -1.559376
MAP marginal 0.24531385 -1.569122

marginal posterior summary:
                tau         mu
mode      0.2453139 -1.5691216
median    0.3445022 -1.5734823
mean      0.3810562 -1.5764366
sd        0.2593672  0.3295298
95% lower 0.0000000 -2.2312306
95% upper 0.8607305 -0.9264079

(quoted intervals are shortest credible intervals.)
```

One can see that the analysis was based on $k = 6$ studies, that both parameters' priors were found to be proper, and maximum-likelihood (ML) as well as maximum-a-posteriori (MAP) values are quoted. Probably most interestingly, under "marginal posterior summary" one can find summary statistics describing the marginal posterior distributions of heterogeneity (τ) and effect (μ), which may often be the most relevant figures. The resulting posterior median and 95% credible interval for the effect μ here are at a log-OR of -1.57 $[-2.23, -0.93]$; this information may eventually constitute the essential result in many cases.

The `forestplot()` *function*

To illustrate data and results, one can use the `forestplot()` function. This function is actually a `bayesmeta`-specific method based on the **forestplot** package's generic `forestplot()` function (Gordon and Lumley 2017). In its simplest form, it may be used as

```
R> forestplot(ma01)
```

Figure 2 shows the `forestplot()` function's default output for the example analysis. In the figure one can see all estimates y_i along with 95% intervals based on the provided standard errors σ_i. At the bottom, 95% credible intervals for the effect and for the predictive distribution are shown (Lewis and Clarke 2001; Guddat *et al.* 2012). Next to each of the quoted estimates (as specified through y_i and σ_i), the *shrinkage intervals* for the study-specific effects θ_i are also

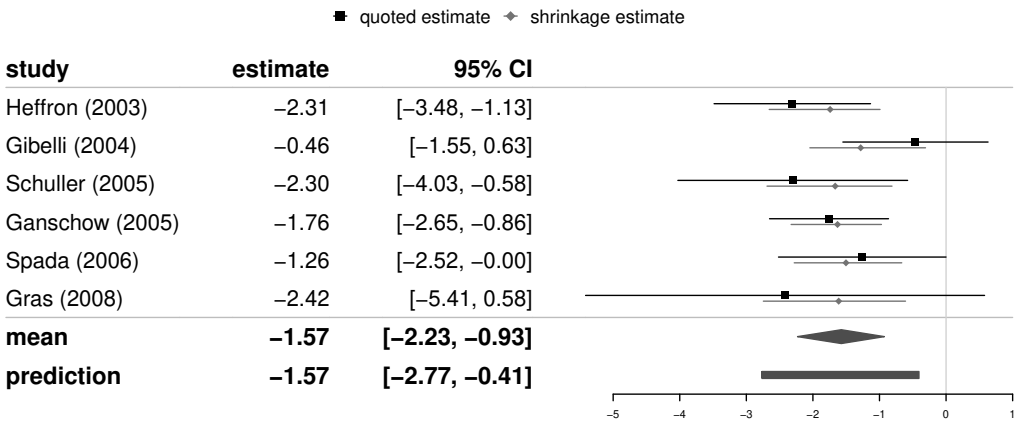

Figure 2: A forest plot, generated using the `forestplot()` function with default settings, showing the input data, effect estimate, prediction interval and shrinkage estimates.

shown in grey; these illustrate the posterior of each individual study's true effect (see equation (2) and Sec. 2.8). The forest plot can be customized in many ways; one can add columns to the table, change axis scaling and labels, omit shrinkage or prediction intervals, etc. For all the options see the online documentation for the `forestplot.bayesmeta()` method.

The `plot()` *function*

The analysis output may be inspected more closely using the `plot()` function:

```
R> plot(ma01)
```

The output for our example is shown in Figure 3; in particular, the joint and marginal posterior distributions are illustrated in detail. Prior densities may be superimposed by using the "`prior=TRUE`" argument, and axis ranges may also be specified manually; see also the online help for the `plot.bayesmeta()` method.

Elements of the `bayesmeta()` *output*

It is possible to access the joint and marginal densities shown in Figure 3 (and more) directly from the `bayesmeta()` output. As usual for an object returned from a non-trivial analysis function, the result of a `bayesmeta()` call is a `list` object of class "`bayesmeta`" containing a number of further individual objects. One can check the complete listing of available entries in the online documentation. For example, there is the "`...$summary`" entry giving some basic summary statistics:

```
R> ma01$summary
```

```
              tau         mu      theta
mode    0.2453139 -1.5691214 -1.5632732
median  0.3445023 -1.5734819 -1.5701653
```

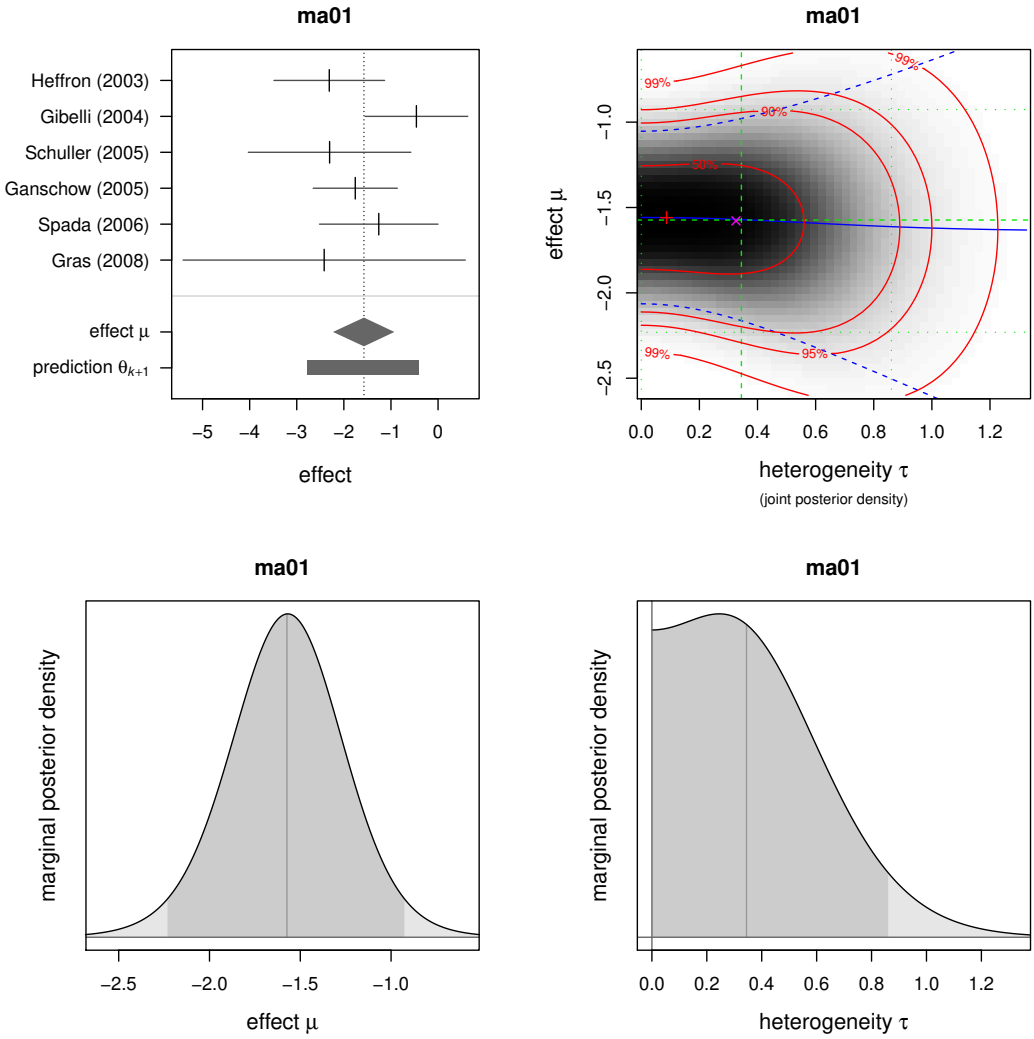

Figure 3: The four plots generated via the `plot()` function. The top left plot is a simple forest plot showing estimates and 95% intervals illustrating the input data (y_i and σ_i) along with the estimated mean effect μ and a prediction interval for the effect θ_{k+1} in a future study. The top right plot illustrates the joint posterior density of heterogeneity τ and effect μ, with darker shading corresponding to higher probability density. The red lines indicate (approximate) 2-dimensional credible regions, and the green lines show marginal posterior medians and 95% credible intervals. The blue lines show the conditional posterior mean effect $\widehat{\mu}(\tau)$ as a function of the heterogeneity τ along with a 95% interval based on its conditional standard error $\widehat{\sigma}(\tau)$ (see also Section 2.4). The red cross (+) indicates the posterior mode, while the pink cross (×) shows the ML estimate. The two bottom plots show the marginal posterior densities of effect μ and heterogeneity τ. 95% credible intervals are indicated with a darker shading, and the posterior median is shown by a vertical line.

```
mean       0.3810562 -1.5764365 -1.5764365
sd         0.2593672  0.3295301  0.5671855
95% lower  0.0000000 -2.2311251 -2.7661319
95% upper  0.8607193 -0.9263075 -0.4072970
```

Some of these we already saw in the output when simply printing the object (see above). The additional third column here shows summary statistics for the predictive distribution of a 'future' study (θ_{k+1}). One can also access the original data (the y_i and σ_i) in the "`...$y`" and "`...$sigma`" entries, or the study labels and the total number of studies (k) in the "`...$labels`" and "`...$k`" entries.

Most importantly, some of the elements are `function`s allowing to access and evaluate the various posterior distributions. For example, the posterior density can be accessed via the "`...$dposterior()`" function; this function has a "`mu`" or a "`tau`" argument, specifying either of these results in a marginal density, and specifying both gives the joint density. So a simple plot of the effect's marginal posterior density can be generated by

```
R> x <- seq(-3, 0.5, length=200)
R> plot(x, ma01$dposterior(mu=x), type="l",
+    xlab="effect", ylab="posterior density")
R> abline(h=0, v=0, col="grey")
```

In order to calculate the posterior probability of a non-beneficial effect ($\mathrm{P}(\mu > 0|\vec{y}, \vec{\sigma}) = 1 - \mathrm{P}(\mu \leq 0|\vec{y}, \vec{\sigma})$), one needs to evaluate the marginal posterior cumulative distribution function (CDF). This is provided via the "`...$pposterior()`" function:

```
R> 1 - ma01$pposterior(mu=0)

[1] 6.187343e-05
```

Or one can also plot the complete CDF using the following code:

```
R> x <- seq(-3, 0.5, length=200)
R> plot(x, ma01$pposterior(mu=x), type="l",
+    xlab="effect", ylab="posterior CDF")
R> abline(h=0:1, v=0, col="grey")
```

The same works also for the heterogeneity parameter τ; in order to derive for example the posterior probability for a "fairly extreme" heterogeneity ($\tau > 1$), one simply needs to supply the "`tau`" parameter instead:

```
R> 1 - ma01$pposterior(tau=1)

[1] 0.02097488
```

so the posterior probability is at 2.1% here. The quantile function (inverse CDF) is also available in the "`...$qposterior()`" function; in order to derive for example a 99% upper limit on the heterogeneity parameter, one needs to evaluate

```
R> ma01$qposterior(tau.p=0.99)

[1] 1.109186
```

so the 99% upper limit would here be at $\tau = 1.11$.

In many cases it is useful to use Monte Carlo simulation to derive other non-trival quantities from the posterior distribution. One can generate samples from the posterior distribution using the "`...$rposterior()`" function. A call of

```
R> ma01$rposterior(n=5)

            tau        mu
[1,] 0.23423926 -1.380271
[2,] 0.28630556 -1.442691
[3,] 0.04402682 -1.610052
[4,] 0.83672662 -1.550758
[5,] 0.18981184 -1.803012
```

will generate a sample of 5 draws from the joint (bivariate) posterior distribution of τ and μ. If one is only interested in the marginal distribution of μ, it is (substantially!) more efficient to omit the τ draws and use

```
R> ma01$rposterior(n=5, tau.sample=FALSE)

[1] -2.184596 -1.876711 -1.514224 -1.384694 -1.567397
```

to generate a vector of μ values only.

For example, suppose that we assume a rate of AR events of $p_c = 50\%$ for the control group, and we are interested in the implied *risk difference* based on our analysis. The risk difference is simply $p_t - p_c$, where p_t is the event rate in the treatment (IL-2RA) group. To determine the distribution of the risk difference we can now simply use Monte Carlo sampling and run

```
R> prob.control <- 0.5
R> logodds.control <- log(prob.control / (1 - prob.control))
R> logodds.treat <- (logodds.control
+                    + ma01$rposterior(n=10000, tau.sample=FALSE))
R> prob.treat <- exp(logodds.treat) / (1 + exp(logodds.treat))
R> riskdiff <- (prob.treat - prob.control)
R> median(riskdiff)

[1] -0.3284975

R> quantile(riskdiff, c(0.025, 0.975))

      2.5%      97.5%
-0.4028368 -0.2149175
```

So here we find a median risk difference of -0.33 and a 95% credible interval of $[-0.40, -0.21]$ for this example. The risk difference distribution could now also be investigated further using histograms etc.

Credible intervals

Central credible intervals can be computed using the corresponding posterior quantiles via the "`...$qposterior()`" function (see above). By default however, *shortest* intervals (see Section 2.9) are provided in the `bayesmeta()` output, or they can also be computed using the `...$post.interval()` function. The `bayesmeta()` function's default behaviour may also be controlled by setting the "`interval.type`" argument. Looking at Figure 3 (marginal posteriors at the bottom), one can see that, depending on the posterior's shape, the shortest intervals may turn out one- or two-sided, at least for the heterogeneity parameter. For example a 99% credible interval for the heterogeneity can then be computed via

```
R> ma01$post.interval(tau.level=0.99)

[1] 0.000000 1.109186
attr(,"interval.type")
[1] "shortest"
```

One can also see that the returned interval contains an attribute indicating the type of interval. A central interval then is derived by explicitly specifying the method to be used for computation:

```
R> ma01$post.interval(tau.level=0.99, method="central")

[1] 0.003547657 1.205400562
attr(,"interval.type")
[1] "central"
```

Such an interval then is actually simply based on the corresponding "central" quantiles, as one may confirm by running:

```
R> ma01$qposterior(tau.p=c(0.005, 0.995))

[1] 0.003547657 1.205400562
```

Prediction

Besides inferring the "main" parameters μ and τ, one can do the same computations for prediction, i.e., a future study's parameter θ_{k+1}. Basic summary statistics for the posterior predictive distribution are already contained in the "`...$summary`" element (see above). The "`...$dposterior()`", "`...$pposterior()`", "`...$qposterior()`", "`...$rposterior()`" and "`...$post.interval()`" functions all have an optional "`predict`" argument to request the predictive distribution. That way, one can for example combine the posterior and predictive densities of μ and θ_{k+1} in a plot:

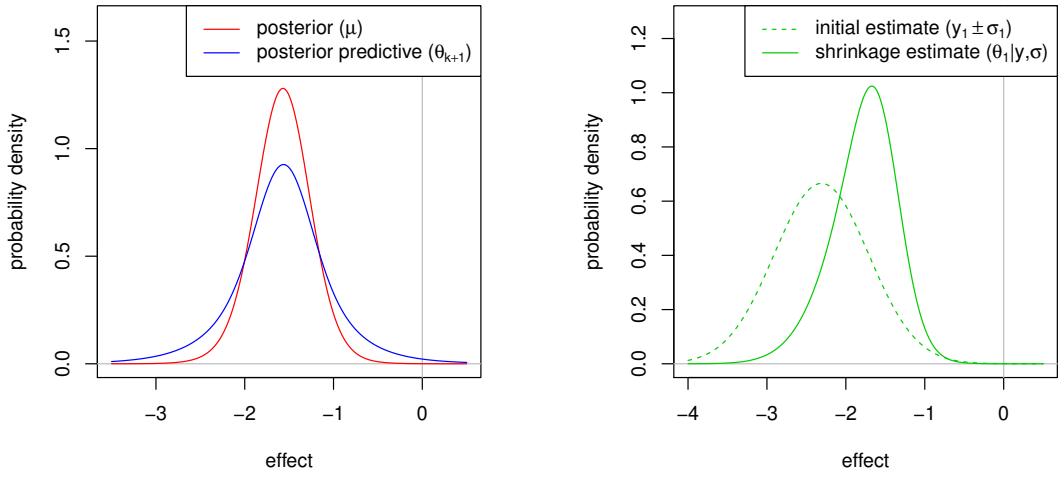

Figure 4: Posterior and posterior predictive densities for the overall effect μ and a 'future' study's parameter θ_{k+1} (left panel), and original (y_1, σ_1) and shrinkage (posterior $\theta_1|\vec{y},\vec{\sigma}$) estimates for the first ($i=1$) study (right panel). The corresponding estimates (medians and 95% credible intervals) are also shown in the forest plot in Figure 2.

```
R> x <- seq(-3.5, 0.5, length=200)
R> plot(x, ma01$dposterior(mu=x), type="n",
+    xlab="effect", ylab="probability density")
R> abline(h=0, v=0, col="grey")
R> lines(x, ma01$dposterior(mu=x), col="red")
R> lines(x, ma01$dposterior(mu=x, predict=TRUE), col="blue")
```

The resulting plot is shown in Figure 4 (left panel). Analogously, the "`predict`" argument may be used to compute e.g. CDFs, quantile functions or credible intervals.

Shrinkage

The "shrinkage" posterior distributions of the study-specific parameters θ_i are also accessible from the `bayesmeta()` output. They are also summarized in the "`...$theta`" element; for example, shrinkage for the first two studies is shown in the first two columns:

```
R> ma01$theta[,1:2]
```

```
          Heffron (2003) Gibelli (2004)
y              -2.3097026     -0.4595323
sigma           0.5994763      0.5563956
mode           -1.6711220     -1.3895876
median         -1.7411356     -1.2821722
mean           -1.7778965     -1.2339736
sd              0.4229425      0.4488759
95% lower      -2.6561049     -2.0455088
95% upper      -0.9920023     -0.3089461
```

One can see the original data (y_i and σ_i) along with the posterior summaries (see also the forest plot in Figure 2). The "`...$dposterior()`", "`...$pposterior()`", "`...$qposterior()`", "`...$rposterior()`" and "`...$post.interval()`" functions again also have an optional "`individual`" argument to specify one of the individual studies (either by their index or by their name). For example, one can illustrate the first study's ($i = 1$) input data (y_1, σ_1) and shrinkage estimate (θ_1) in a single plot using the following code

```
R> x <- seq(-4, 0.5, length=200)
R> plot(x, ma01$dposterior(theta=x, individual=1), type="n",
+    xlab="effect", ylab="probability density")
R> abline(h=0, v=0, col="grey")
R> lines(x, dnorm(x, mean=ma01$y[1], sd=ma01$sigma[1]),
+    col="green", lty="dashed")
R> lines(x, ma01$dposterior(theta=x, individual=1), col="green")
```

The resulting two densities are shown in Figure 4 (right panel). Analogously, the "`individual`" argument may be used to compute e.g. CDFs, quantile functions or credible intervals.

3.4. Investigating prior variations

Prior predictive distributions

In order to judge the implications of settings of the heterogeneity prior, it is often useful to consider *prior predictive distributions* (Gelman *et al.* 2014). Any fixed value of τ will imply a certain (prior) distribution $p(\theta_i|\mu, \tau)$ and variability among the true study-specific means $\theta_1, \ldots, \theta_k$, namely, a normal distribution with $\text{Var}(\theta_i|\mu, \tau) = \tau^2$ (see also (2)). Depending on the type of endpoint (e.g., log-ORs), the implied variability can be interpreted and judged on the corresponding outcome scale (Spiegelhalter *et al.* 2004, Sec. 5.7).

Assuming a prior distribution for τ, rather than a fixed value, also implies assumptions on the a priori expected distribution and variability of the true study parameters θ_i. The *prior predictive* distribution $p(\theta_i|\tau)$ of the θ_i values then is a mixture of normal distributions, with mean μ and with the prior $p(\tau)$ as the mixing distribution for the normal standard deviation (Seidel 2010; Lindsay 1995). As the name suggests, the prior predictive distribution is actually closely related to the (posterior) predictive distribution discussed above (Gelman *et al.* 2014). This mixture distribution can again be evaluated using the DIRECT algorithm (Röver and Friede 2017); this approach is implemented in the `normalmixture()` function.

Consider the half-normal prior distribution with scale 0.5 that was used for the heterogeneity in the above analysis. We can now check what prior predictive distribution this prior corresponds to. We only need to supply the prior CDF (the mean μ is by default set to zero):

```
R> hn05 <- normalmixture(cdf=function(t){phalfnormal(t, scale=0.5)})
```

One can check the returned result (e.g. via `str(hn05)`); the result is a `list` with several elements, among which most importantly are the mixture's density, cumulative distribution and quantile functions ("`...$density()`", "`...$cdf()`" and "`...$quantile()`", respectively).

For comparison, we can also check the implications of a half-Cauchy prior of the same scale, or a half-normal prior of doubled scale:

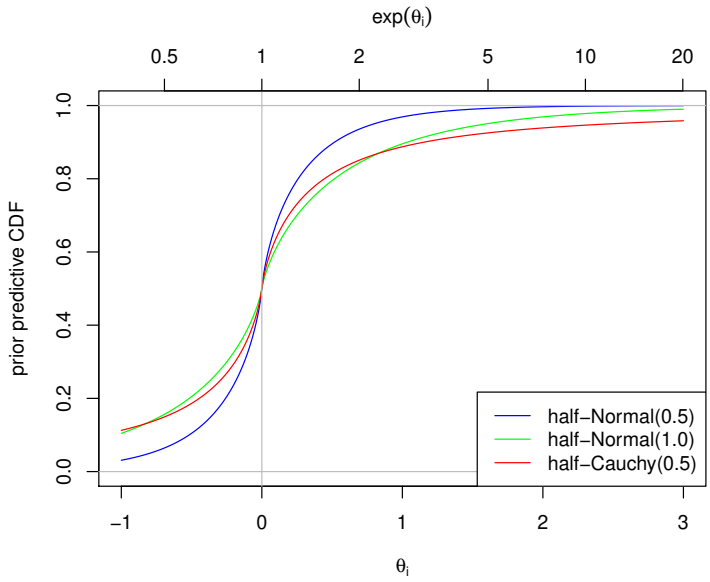

Figure 5: Prior predictive distributions for the true study means θ_i assuming several heterogeneity priors (and $\mu = 0$) as computed using the `normalmixture()` function.

```
R> hn10 <- normalmixture(cdf=function(t){phalfnormal(t, scale=1.0)})
R> hc05 <- normalmixture(cdf=function(t){phalfcauchy(t, scale=0.5)})
```

and compare these graphically via their implied prior predictive CDFs by accessing the three mixtures' "`...$cdf()`" functions:

```
R> x <- seq(-1, 3, by=0.01)
R> plot(x, hn05$cdf(x), type="l", col="blue", ylim=0:1,
+    xlab=expression(theta[i]), ylab="prior predictive CDF")
R> lines(x, hn10$cdf(x), col="green")
R> lines(x, hc05$cdf(x), col="red")
R> abline(h=0:1, v=0, col="grey")
R> axis(3, at=log(c(0.5,1,2,5,10,20)), lab=c(0.5,1,2,5,10,20))
R> mtext(expression(exp(theta[i])), side=3, line=2.5)
```

The resulting plot is shown in Figure 5. In our example, the effect measure is a log-OR, so the θ_i need to be interpreted on the exponentiated scale (see the top axis). A priori, 95% of θ_i values are assumed to be within $\pm$ the 97.5% quantile of the (symmetric) prior predictive distribution. We can now check what this means for our three cases:

```
R> q975 <- c("half-normal(0.5)" = hn05$quantile(0.975),
+            "half-normal(1.0)" = hn10$quantile(0.975),
+            "half-Cauchy(0.5)" = hc05$quantile(0.975))
R> print(cbind("theta"=q975, "exp(theta)"=exp(q975)))
```

```
                      theta exp(theta)
half-normal(0.5) 1.092287   2.981083
half-normal(1.0) 2.184573   8.886857
half-Cauchy(0.5) 5.050571 156.111517
```

So for the half-normal prior with scale 0.5 we have 95% probability roughly within a factor of $\frac{1}{3}$ or 3 around the overall mean odds ratio ($\exp(\mu)$). For the other two priors, the numbers are much more extreme.

Informative heterogeneity priors

It may also make sense to consider empirical information for the setup of an informative heterogeneity prior, for example, when other evidence is extremely sparse. In medical or psychological contexts, some evidence for certain types of endpoints may be found e.g. in Pullenayegum (2011), Turner *et al.* (2012), Kontopantelis *et al.* (2013) and van Erp *et al.* (2017). Instantly applicable for a meta-analysis are the numbers given by Rhodes *et al.* (2015) and Turner *et al.* (2015), where in both cases the complete *Cochrane Database of Systematic Reviews* was analyzed to infer the predictive distribution of heterogeneity for specific applications. The investigation by Rhodes *et al.* (2015) here was concerned with mean difference endpoints, while Turner *et al.* (2015) focused on log-OR endpoints. The derived prior distributions are directly available in the **bayesmeta** package via the `RhodesEtAlPrior()` and `TurnerEtAlPrior()` functions. For our present example (a log-OR endpoint whose definition may be categorized as "surgical / device related success / failure", and where the comparison is between pharmacological treatment and control), we can derive the prior simply as

```
R> tp <- TurnerEtAlPrior(outcome = "surgical",
+    comparator1 = "pharmacological", comparator2 = "placebo / control")
```

For a complete description of the possible input options see the online documentation; the `RhodesEtAlPrior()` function then works similarly. The function output is a `list` with several entries, including the prior density, cumulative distribution and quantile function (in this case a log-normal distribution) in the "`...$dprior()`", "`...$pprior()`" and "`...$qprior()`" elements. This way we can e.g. check what magnitude of heterogeneity values is a priori expected for this setting by determining the median as well as 2.5% and 97.5% quantiles:

```
R> tp$qprior(c(0.025, 0.5, 0.975))

[1] 0.06233896 0.34300852 1.88734045
```

The prior density can now immediately be used and passed on to the `bayesmeta()` function; for example, we can use the same effect prior as before and the "empirical" prior for the heterogeneity:

```
R> ma02 <- bayesmeta(crins.es, mu.prior.mean = 0, mu.prior.sd = 4,
+    tau.prior = tp$dprior)
```

Comparing the results to the previous analysis (e.g. via their "`...$summary`" outputs), one can see that in this case they are very similar. The two corresponding prior densities are also shown in Figure 1 (page 12; solid and dotted blue lines).

Non-informative priors

As discussed in Section 2.2, an obvious choice of an uninformative prior for the effect μ would be the (improper) uniform prior on the real line; this one can be utilized by simply leaving the `mu.prior.mean` and `mu.prior.sd` parameters unspecified. In order to use one of the uninformative heterogeneity priors discussed in Section 2.2, these do not need to be specified "manually" in terms of their probability density function; a set of priors is already pre-implemented and may be specified via a character string. The default setting for example is `tau.prior="uniform"`. If one wants to use, say, the uniform effect prior along with the Jeffreys prior for the heterogeneity τ (see Section 2.2 and Bodnar *et al.* (2017)), one can run

```
R> ma03 <- bayesmeta(crins.es, tau.prior="Jeffreys")
```

The complete list of possible options is described in detail in the online documentation.

3.5. Making the connection with frequentist results

Frequentist and Bayesian approaches to inference within the NNHM framework are obviously related, and it may be interesting to highlight the connection between the corresponding results. A simple frequentist analysis may be performed e.g. using the **metafor** package's "`rma()`" function via

```
R> ma04 <- rma(crins.es)
```

(Viechtbauer 2010). By default, the *restricted ML* (REML) heterogeneity estimator $\hat{\tau}_{\mathrm{REML}}$ is used, but the exact type of estimator does not matter here. The heterogeneity point estimate here turns out as:

```
R> sqrt(ma04$tau2)

[1] 0.4670268
```

and we can retrieve the effect estimate and its standard error via:

```
R> ma04$b

             [,1]
intrcpt -1.591513

R> ma04$se

[1] 0.3340882
```

In the Bayesian setup, these numbers correspond to conditional posterior moments of the effect $(\mu|\tau=\hat{\tau}_{\mathrm{REML}})$ in an analysis using the uniform effect prior. Such an analysis was performed in the previous section (uniform effect prior and Jeffreys heterogeneity prior; the heterogeneity prior does not matter here) and stored in the "`ma03`" object. From this we can retrieve the effect's conditional posterior moments (mean and standard deviation for $\tau=\hat{\tau}_{\mathrm{REML}}$) using the "`...$cond.moment()`" function:

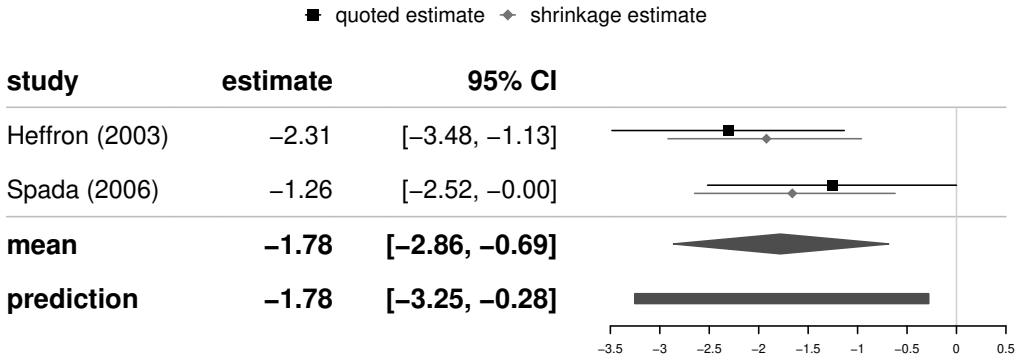

Figure 6: Forest plot showing the data and derived estimates for the analysis of the two randomized studies only.

```
R> ma03$cond.moment(tau = sqrt(ma04$tau2))

          mean        sd
[1,] -1.591513 0.3340882
```

and we can see that these correspond exactly to the frequentist effect estimate. Both analysis approaches are related through the use of the same likelihood function; in the Bayesian analysis uncertainty (e.g. in the heterogeneity) is accounted for via integration, and a prior distribution for both parameters is considered.

3.6. Posterior predictive checks

A meta-analysis of two studies

Posterior predictive p-values allow to quantify the consistency of the data with certain parametric hypotheses; see Section 2.10. In the following we will determine some p-values from the `bayesmeta()` output; to this end, we will investigate a second meta-analysis example involving only two studies.

Of the six studies considered in the pediatric transplantation example (see Figure 2), only two were randomized (Heffron *et al.* 2003; Spada *et al.* 2006). Since randomized studies are usually considered as evidence of higher quality, now suppose one was interested in combining the randomized studies only. Computations analogous to the preceding example may be done via

```
R> ma05 <- bayesmeta(crins.es[crins.es[,"randomized"]=="yes",],
+    mu.prior.mean=0, mu.prior.sd=4,
+    tau.prior=function(t){dhalfnormal(t,scale=0.5)})
```

Figure 6 shows the forest plot for this analysis. Based on these two studies only, we can now inspect e.g. the estimate of the overall effect μ; comparing to the previous analysis (Figure 2), the (absolute) estimate is slightly larger, but the credible interval is wider.

Posterior predictive p-values for the effect (μ)

The obvious 'null' hypothesis to be tested here is $H_0 : \mu \geq 0$ (i.e., no effect or a harmful effect) versus the alternative $H_1 : \mu < 0$ (a beneficial effect). We may now derive a posterior predictive p-value in order to express to what extent the data are consistent with or in contradiction to the null hypothesis. In order to evaluate the "discrepancy" between data and null hypothesis, we need a *test statistic* or *discrepancy variable* that in some sense measures or captures this (in-) compatibility.

An obvious candidate may e.g. be the posterior probability of a beneficial effect, $\mathrm{P}(\mu < 0 \,|\, y)$. This probability here is identical to the posterior cumulative distribution function (CDF) evaluated at the hypothesized value $\mu = 0$. Large values then are evidence *against*, and small values speak *in favour of* the null hypothesis. In the present example data set we can evaluate this figure as

```
R> ma05$pposterior(mu=0)

[1] 0.9974968
```

Regarding our hypothesis setup, the question then is, how (un-) likely our observed value of 0.9975 is under the null hypothesis ($H_0 : \mu \geq 0$). In order to answer that question, we need the posterior distribution of the test statistic conditional on the null hypothesis (and the data). Using Monte Carlo sampling, we can generate draws of parameters from the conditional posterior distribution $(\mu^\star, \tau^\star, \theta^\star \,|\, y, \mu \geq 0)$ and then generate new data based on these $(y^\star \,|\, \mu^\star, \tau^\star, \theta^\star)$ from which we can compute replications of the test statistic and determine its distribution.

In the **bayesmeta** package, posterior predictive checks are implemented in the `pppvalue()` function. In order to generate posterior predictive draws, we need to specify the involved hypotheses, the test statistic, and the number of Monte Carlo replications to be generated; here we use $n = 1000$, which may take a few minutes to compute:

```
R> p1 <- pppvalue(ma05, parameter="mu", value=0, alternative="less",
+    statistic="cdf", n=1000)
```

Since the p-value is eventually computed based on the generated Monte Carlo samples, a value of $n \gg 100$ will usually be appropriate. By default, a progress bar is shown during computation, allowing to estimate the remaining computation time. We can then inspect the result by printing the returned object:

```
R> p1

        'bayesmeta' posterior predictive p-value (one-sided)

data:  ma05
cdf = 0.9975, Monte Carlo replicates = 1000, p-value = 0.01
alternative hypothesis: true effect (mu) is less than 0
```

The default output restates the hypothesis setup and shows a posterior predictive p-value of 0.01. This means that in 10 of the 1000 replications generated (1%) the statistic was larger

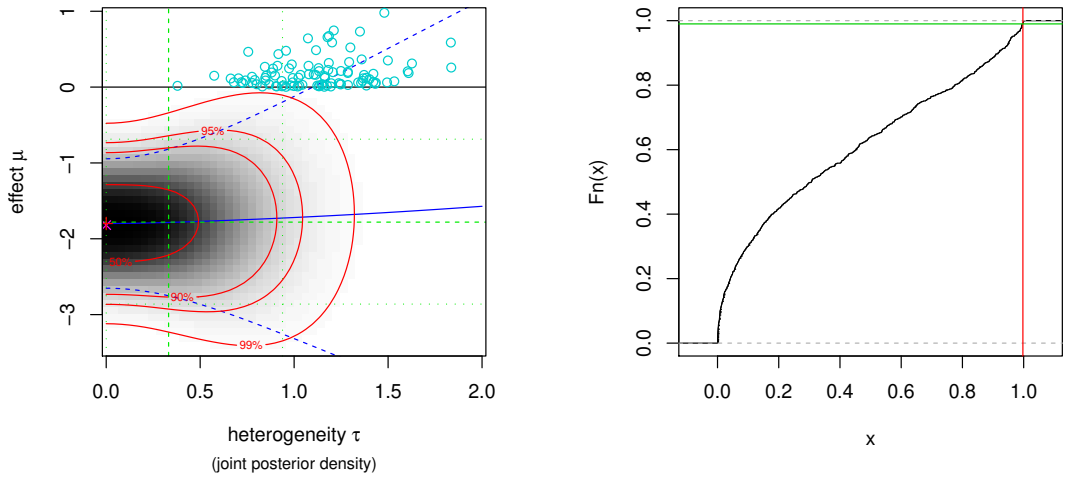

Figure 7: Illustration of the computation of a posterior predictive p-value using Monte Carlo sampling. The left panel shows the distribution of replicated $\mu^\star$ and $\tau^\star$ values, the right panel shows the empirical (cumulative) distribution of the associated "test statistic" values.

than our observed 0.9975. One can also do a quick check of the uncertainty in this Monte-Carlo'ed p-value using e.g. the `prop.test()` function, which here yields a 95% confidence interval ranging roughly from 0.5% to 2%.

The replications are also stored in detail in the generated object (here: "`p1`"). The `list` object contains a "`...$replicates`" element, which again contains vectors of generated $\tau^\star$ and $\mu^\star$ draws, matrices of the corresponding $\theta^\star$ and $y^\star$ draws, and finally the test statistic values along with an indicator showing which ones constitute the "tail area" the p-value is based on. Using the provided output, one can visualize how the posterior predictive p-value is computed; executing

```
R> plot(ma05, which=2, mulim=c(-3.5, 1), taulim=c(0,2))
R> abline(h=p1$null.value)                    # (the null-hypothesized mu value)
R> points(p1$replicates$tau, p1$replicates$mu, col="cyan")    # (the samples)
```

one can see the joint posterior distribution of heterogeneity τ and effect μ along with the generated samples, which, according to the specified null hypothesis, are confined to $\mu \geq 0$ (see Figure 7, left panel). The resulting test statistic values can be illustrated via their empirical cumulative distribution function, which can be generated by

```
R> plot(ecdf(p1$replicates$statistic[,1]))
R> abline(v = p1$statistic, col="red")
R> abline(h = 1-p1$p.value, col="green")
```

(see Figure 7, right panel). The "test statistic" values range between 0 and 1, and their distribution is clearly not uniform. The actualized value in the present data set (0.9975, vertical red line) is situated in the upper tail of the distribution of replicated statistics values, and the remaining tail area (horizontal green line) eventually defines the p-value.

Posterior predictive p-values for the heterogeneity (τ)

Computation of posterior predictive p-values for the heterogeneity works analogously. Use of the posterior CDF ($\mathrm{P}(\tau \leq 0|y)$) to test for zero heterogeneity does not make sense, as this figure will always be zero, for the original as well as any replicated data. In order to test for zero heterogeneity, we could use the classical Cochran's Q statistic:

```
R> p2 <- pppvalue(ma05, parameter="tau", value=0, alternative="greater",
+    statistic="q", n=1000)
```

which here yields a p-value of 24.4%. In this case computations are much faster, since computationally expensive re-analyses of the data are not necessary to compute the test statistic. The resulting p-value should be identical to the "classical" result, since under the null hypothesis considered (here: $\tau = 0$) the Q-statistic follows a χ^2-distribution, as in the frequentist setting.

In a Bayesian context, it may also make sense to consider using for example the Bayes factor of the hypothesis of $\tau = 0$ as the "test statistic" or "discrepancy measure". The `pppvalue()` function is able to utilize arbitrary functions as a statistic; to use the Bayes factor, we can define the function

```
R> BF <- function(y, sigma)
+  {
+    bm <- bayesmeta(y=y, sigma=sigma,
+      mu.prior.mean=0, mu.prior.sd=4,
+      tau.prior=function(t){dhalfnormal(t, scale=0.5)},
+      interval.type="central")
+    return(bm$bayesfactor[1,"tau=0"])
+  }
```

Two things are worth noting here. Firstly, it makes sense to use matching (especially prior) specifications for the `bayesmeta()` call within the `BF()` function as for the original analysis (here: the previously generated "`ma05`" object). Secondly, the use of central intervals (see the "`interval.type`" argument) is more efficient, since these are faster to compute, and the intervals are otherwise irrelevant here. In order to utilize the function for a posterior predictive p-value, we can then call

```
R> p3 <- pppvalue(ma05, parameter="tau", value=0, alternative="greater",
+    statistic=BF, rejection.region="lower.tail", n=1000, sigma=ma05$sigma)
```

Note that the rejection region needs to be specified explicitly here (small Bayes factors constitute evidence *against* the null hypothesis). Additional arguments may be passed to the `statistic` function, like the "`sigma`" argument above. The Bayes factor in this case yields a similar p-value to Cochran's Q statistic ($p = 22.2\%$).

Posterior predictive p-values for individual effects (θ_i)

Quite commonly in a meta-analysis, interest may also be in one of the study specific parameters θ_i (Schmidli *et al.* 2014; Wandel *et al.* 2017). For example, suppose that at the end of

the latter of the two concerned studies (Spada, 2006) a meta-analysis was performed to evaluate the cumulative evidence, but main interest still was in the outcome of the second study that had just been conducted; it would then only be considered in the light of the previous evidence. In such a scenario, we can then evaluate a posterior predictive p-value for the 2nd study's effect (θ_2); this shrinkage estimate is also shown in Figure 6. Using the `pppvalue()` function, we can simply refer to a particular study's parameter by its index or its label:

```
R> p4 <- pppvalue(ma05, parameter="Spada", value=0, alternative="less",
+    statistic="cdf", n=1000)
```

which here results in a p-value of around 16.1%.

4. Summary

A Bayesian approach has distinct advantages in the context of meta-analysis; it allows to coherently process the uncertainty in the heterogeneity (nuisance) parameter while focusing on inference for the effect parameter(s), small sample sizes (numbers of studies) do not pose a difficulty, and interpretation of the results is very straightforward. Since meta-analyses are quite commonly based on only very few studies, the opportunity to formally utilize external information in the analysis via the prior specification may be a welcome feature. Unlike for some other methods whose results depend on the specification of secondary details, a Bayesian analysis result is uniquely defined once the model (likelihood and prior) is specified.

The application of Bayesian reasoning for this purpose is not a novelty (Spiegelhalter *et al.* 2004), but it usually comes with a certain computational burden; often MCMC methods are necessary, which demand a substantial amount of attention on their own (Gilks *et al.* 1996). The **bayesmeta** package (Röver 2015) allows to perform Bayesian random-effects meta-analyses without the need to worry too much about the computational details. Some of the technical details of the computational approach underlying the package have been described elsewhere (Röver and Friede 2017). The simple normal-normal hierarchical model (NNHM) treated here is applicable in a wide range of contexts and is routinely used for many types of input data and effect measures (Hedges and Olkin 1985; Hartung *et al.* 2008; Viechtbauer 2010; Borenstein *et al.* 2009). The **bayesmeta** implementation allows for quick, accurate and reproducible computation, and it has already faciltated some larger-scale simulation studies to compare Bayesian results with common alternative approaches and evaluate their relative performance (Friede *et al.* 2017a,b). Usage of the **bayesmeta** package is not more complicated to use than many other common meta-analysis tools. The availability of predictive distributions and shrinkage estimates makes the package attractive also for advanced evidence synthesis applications, like extrapolation to future studies (Schmidli *et al.* 2014, 2017; Wandel *et al.* 2017). Since the generic NNHM appears in different fields of application, use of the **bayesmeta** package may also be extended to other areas of research beyond common meta-analysis. For example, it could as well be used to model hierarchical structures *within* a study (e.g., groups of patients), or a two-stage approach may be useful for meta-analysis based on individual-patient data. In future, the same numerical approach might be extended to the more general case of meta-regression.

A. Appendix

A.1. Unit information priors for binary outcomes

Logarithmic odds ratios (log-OR)

If the effect measure of an analysis is a logarithmic odds ratio (log-OR; see Section 3.2), then the standard error derived from a 2×2 contingency table amounts to $\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$, where a, b, c and d are the four entries (counts) and $N=a+b+c+d$ is the total number of subjects (Hedges and Olkin 1985; Hartung *et al.* 2008; Borenstein *et al.* 2009). Assuming equal allocation and a neutral effect, we can simply set the table allocation to $a=b=c=d=\frac{N}{4}$. If we further plug in a total sample size of $N=1$, this leads (heuristically) to a unit information prior for the log-OR with zero mean and a standard deviation of 4. For this prior distribution, log-ORs are within a range of ±7.84 with 95% probability, corresponding to ORs roughly within a range from $\frac{1}{2500}$ to 2500.

If more generally we consider the case of a particular event probability $p\in[0,1]$, we can derive a unit information prior by assuming $a=c=p\frac{N}{2}$ and $b=d=(1-p)\frac{N}{2}$, leading to a generally even larger prior standard deviation of $\frac{2}{\sqrt{p(1-p)}}$.

Logarithmic relative risks (log-RR)

Similarly to the previous section, the logarithmic relative risk (log-RR) is given by $\log\left(\frac{a/(a+b)}{c/(c+d)}\right)$, and its associated standard error is $\sqrt{\frac{1}{a}-\frac{1}{a+b}+\frac{1}{c}-\frac{1}{c+d}}$ (Hedges and Olkin 1985; Hartung *et al.* 2008; Borenstein *et al.* 2009). Again plugging in $a=b=c=d=\frac{N}{4}$, this now amounts to a standard deviation of 2. If we introduce a certain event probability p (and plugging in $a=c=p\frac{N}{2}$ and $a+b=c+d=\frac{N}{2}$, the error is $2\sqrt{\frac{1-p}{p}}$, which is larger for $p<\frac{1}{2}$ and smaller for $p>\frac{1}{2}$.

A.2. Conservatism of the uniform heterogeity prior

As discussed in Section 2.2, it is hard to define an "uninformative" prior for the heterogeneity parameter τ. A larger heterogeneity will first of all generally lead to a larger variance of the effect's marginal posterior (via the larger variance of the *conditional* distribution; see equations (17), (21)). One may then argue that an overestimation of heterogeneity may be considered a *conservative* form of bias, so that, for example, among two exponential prior distributions the one with the larger expectation was "more conservative" in a certain sense.

A shift in heterogeneity causes a change in *both* the conditional standard deviation *and* mean. If the shift in $\widehat{\mu}(\tau)$ happens to be larger than the shift in $\widehat{\sigma}(\tau)$, then the resulting (conditional) confidence interval for a larger heterogeneity value does not necessarily completely contain the interval corresponding to a smaller heterogeneity. Such cases may then lead to counterintuitive results for (frequentist) fixed- and random-effects analysis results especially in settings with imbalanced standard errors (Poole and Greenland 1999). Although it is not obvious whether such pathologies are also realistic in a Bayesian analysis, *a priori*, this is unlikely to lead to any systematic bias.

Nevertheless, along these lines it is possible to show a particular "conservatism" property for

the improper uniform prior. The derivation goes as follows. Suppose we have a bounded parameter domain $[a, \infty]$, a likelihood function $f(x) \geq 0$ ($x \in [a, \infty]$) with $\int_a^\infty f(x)\,\mathrm{d}x < \infty$, and a prior with monotonically decreasing probability density function $p(\cdot)$, so that $0 < p(a) < \infty$, and $a \leq x < y \;\Rightarrow\; 0 \leq p(y) \leq p(x)$. Using the (improper) uniform prior or prior p we get different posteriors with cumulative distribution functions $F_1(\cdot)$ and $F_p(\cdot)$, respectively. From the above assumptions follows that

$$F_p(y) = \frac{\int_a^y f(x)\,p(x)\,\mathrm{d}x}{\int_a^\infty f(x)\,p(x)\,\mathrm{d}x} \geq \frac{\int_a^y f(x)\,p(y)\,\mathrm{d}x}{\int_a^\infty f(x)\,p(a)\,\mathrm{d}x} = \frac{p(y)}{p(a)}\,\frac{\int_a^y f(x)\,\mathrm{d}x}{\int_a^\infty f(x)\,\mathrm{d}x} \tag{30}$$

$$\geq \frac{\int_a^y f(x)\,\mathrm{d}x}{\int_a^\infty f(x)\,\mathrm{d}x} = F_1(y) \tag{31}$$

for all $y > a$. This means that with $F_p(y) \geq F_1(y)$, the posterior using the uniform prior is *stochastically larger* than the posterior based on any other prior among the class of priors with monotonically decreasing density $p(\cdot)$ and finite $p(a)$ (provided the uniform prior yields a proper posterior). In our context, this especially implies that quantiles or expectations based on the uniform prior are larger (Shaked and Shanthikumar 2007).

The class of priors with finite intercept and monotonically decreasing density to which the above property applies includes e.g. the exponential, half-normal, half-Student-t, half-Cauchy and Lomax distributions (Johnson *et al.* 1994), or uniform distributions with a finite upper bound.

A.3. Marginal likelihood derivation

Using the improper uniform prior for μ ($p(\mu) \propto 1$), the marginal likelihood, marginalized over μ, is

$$p(\vec{y}|\tau, \vec{\sigma}) = \int p(\vec{y}|\mu, \tau, \vec{\sigma})\,p(\mu)\,\mathrm{d}\mu \tag{32}$$

$$= (2\pi)^{-\frac{k}{2}} \times \prod_{i=1}^{k} \frac{1}{\sqrt{\sigma_i^2+\tau^2}} \times \int \exp\left(-\frac{1}{2}\sum_{i=1}^{k}\frac{(y_i-\mu)^2}{\sigma_i^2+\tau^2}\right)\mathrm{d}\mu \tag{33}$$

where

$$\sum_{i=1}^{k}\frac{(y_i-\mu)^2}{\sigma_i^2+\tau^2} = \underbrace{\sum_{i=1}^{k}\frac{y_i^2}{\sigma_i^2+\tau^2}}_{=:a} + \mu\underbrace{\left(-2\sum_{i=1}^{k}\frac{y_i}{\sigma_i^2+\tau^2}\right)}_{=:b} + \mu^2\underbrace{\sum_{i=1}^{k}\frac{1}{\sigma_i^2+\tau^2}}_{=:c} \tag{34}$$

$$= a + b\,\mu + c\,\mu^2 \tag{35}$$

$$= \frac{(\mu - \frac{-b}{2c})^2}{\sqrt{1/c}^2} + a - \frac{b^2}{4c} = \frac{(\mu - \widehat{\mu}(\tau))^2}{\widehat{\sigma}(\tau)^2} + \Delta(\tau) \tag{36}$$

and

$$\widehat{\mu}(\tau) = \frac{-b}{2c} = \frac{\sum_{i=1}^{k}\frac{y_i}{\sigma_i^2+\tau^2}}{\sum_{i=1}^{k}\frac{1}{\sigma_i^2+\tau^2}} \tag{37}$$

$$\widehat{\sigma}(\tau) = \sqrt{1/c} = \sqrt{\frac{1}{\sum_{i=1}^{k}\frac{1}{\sigma_i^2+\tau^2}}} \tag{38}$$

$$\Delta(\tau) \;=\; a - \tfrac{b^2}{4c} \;=\; \sum_{i=1}^{k} \frac{y_i^2}{\sigma_i^2+\tau^2} - \frac{\left(\sum_{i=1}^{k} \frac{y_i}{\sigma_i^2+\tau^2}\right)^2}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2+\tau^2}}. \tag{39}$$

$$= \sum_{i=1}^{k} \frac{1}{\sigma_i^2+\tau^2}\left(y_i - \sum_{j=1}^{k} \frac{\frac{1}{\sigma_j^2+\tau^2}\, y_j}{\sum_{\ell=1}^{k} \frac{1}{\sigma_\ell^2+\tau^2}}\right)^2 = \sum_{i=1}^{k} \frac{\left(y_i - \widehat{\mu}(\tau)\right)^2}{\sigma_i^2+\tau^2}. \tag{40}$$

Note that $\widehat{\mu}(\tau)$ (37) and $\widehat{\sigma}(\tau)$ (38) are the conditional posterior mean and standard deviation of $\mu|\tau$. With that the marginal likelihood turns out as

$$p(\vec{y}|\tau,\vec{\sigma}) \;=\; (2\pi)^{-\frac{k}{2}} \times \prod_{i=1}^{k} \frac{1}{\sqrt{\sigma_i^2+\tau^2}} \times \int \exp\left(-\frac{1}{2}\frac{\left(\mu-\widehat{\mu}(\tau)\right)^2}{\widehat{\sigma}^2(\tau)} - \frac{1}{2}\Delta(\tau)\right) \mathrm{d}\mu \tag{41}$$

$$= (2\pi)^{-\frac{k}{2}} \times \prod_{i=1}^{k} \frac{1}{\sqrt{\sigma_i^2+\tau^2}} \times \exp(-\tfrac{1}{2}\Delta(\tau)) \times \int \exp\left(-\frac{1}{2}\frac{\left(\mu-\widehat{\mu}(\tau)\right)^2}{\widehat{\sigma}^2(\tau)}\right) \mathrm{d}\mu \tag{42}$$

$$= (2\pi)^{-\frac{k}{2}} \times \prod_{i=1}^{k} \frac{1}{\sqrt{\sigma_i^2+\tau^2}} \times \exp(-\tfrac{1}{2}\Delta(\tau)) \times \sqrt{2\pi}\,\widehat{\sigma}(\tau) \tag{43}$$

$$= (2\pi)^{-\frac{k-1}{2}} \times \prod_{i=1}^{k} \frac{1}{\sqrt{\sigma_i^2+\tau^2}} \times \exp\left(-\frac{1}{2}\frac{\left(y_i-\widehat{\mu}(\tau)\right)^2}{\sigma_i^2+\tau^2}\right) \times \frac{1}{\sqrt{\sum_{i=1}^{k} \frac{1}{\sigma_i^2+\tau^2}}}. \tag{44}$$

The derivation for an informative normal effect prior (with mean μ_p and variance σ_p^2) works similarly.

A.4. Mixture implementation details

The approximation of marginal effect distributions etc. is implemented via the DIRECT algorithm as described by Röver and Friede (2017). This approximation is utilized to evaluate posterior distributions of the overall effect μ as well as shrinkage estimates θ_i and predictions θ_{k+1}. In all three cases, the distributions of interest are mixtures of conditionally normal distributions; in order to construct the approximate discrete mixture, it is necessary to evaluate *symmetrized divergences* of the conditional distributions. The symmetrized divergence (relative entropy) for two normal distributions with mean and variance parameters (μ_A, σ_A^2) and (μ_B, σ_B^2), respectively, is given by

$$\mathcal{D}_\mathrm{s}(p(\vartheta|\mu_A,\sigma_A)\|p(\vartheta|\mu_B,\sigma_B)) \;=\; \frac{(\mu_A-\mu_B)^2}{\left(\frac{1}{2}(\sigma_A^{-2}+\sigma_B^{-2})\right)^{-1}} + \frac{(\sigma_A^2-\sigma_B^2)^2}{2\,\sigma_A^2\,\sigma_B^2} \tag{45}$$

(Röver and Friede 2017). Since the conditional means of $\mu|\tau$ and $\theta_{k+1}|\tau$ are identical, while the conditional variance of the latter is always equal to or larger than the former (see Section 2.8), a grid constructed for the effect's posterior (μ) can always be used for the predictive distribution (θ_{k+1}) without loss of accuracy. In order not to have to construct and consider several separate τ grids also for the different shrinkage distributions, the general algorithm is slightly extended. Instead of determining divergences corresponding to pairs of τ values with respect to *each* of the shrinkage distributions individually, the *maximum* divergence across effect posterior as well as all k shrinkage distributions is considered. The result is a single grid in τ values that may be re-used for all three types of distributions.

A.5. Calibration check

The inferential statements returned by a Bayesian analysis differ in their probabilistic claims from those returned by frequentist analyses. For example, while a frequentist 95% confidence interval usually is supposed to yield 95% coverage for repeated data generation and analysis *conditional on any single point in parameter space*, a Bayesian analysis is to be understood *conditional on the assumed prior distribution*, and hence the coverage holds for repeated sampling of parameters from the prior and subsequent data generation and analysis. While frequentist analyses often rely on large-sample-size asymptotics (here: large k), Bayesian posterior analyses generally should (at least for proper priors) yield exact coverages, independent of sample sizes (Dawid 1982; Gneiting *et al.* 2007). The accuracy (calibration) of Bayesian analysis software may be checked exploiting this property (Cook *et al.* 2006). The aim of this section is to demonstrate that the **bayesmeta** implementation in fact yields consistent results.

If a Bayesian analysis method is properly calibrated, then the repeated subsequent generation of (i) parameter values $\theta^\star$ from the prior distribution $p(\theta)$, (ii) data $y^\star$ from the conditional sampling distribution $p(y|\theta^\star)$, and (iii) posterior probabilities $p^\star = \mathrm{P}(\theta \leq \theta^\star|y^\star)$ will yield a sample of so-called *probability integral transform (PIT)* values $p^\star$ (Gneiting *et al.* 2007). If the implementation is accurate, then these PIT values follow a uniform probability distribution. Investigation of the empirical cumulative distribution function(s) of individual parameters' PIT values returns the empirical frequency with which a one-sided credible interval of a given credible level would have covered the true value across the generated parameter and data samples. This way it allows to investigate the fidelity of the analysis procedure across the prior's domain as well as across credible levels (Cook *et al.* 2006). For the meta-analysis problem within the NNHM, such a calibration check may be implemented as follows:

```
R> mupriormean   <- 0.0
R> mupriorsd     <- 4.0
R> taupriorscale <- 0.5
R> Nsim <- 1000
R> pit <- matrix(NA, nrow=Nsim, ncol=2, dimnames=list(NULL, c("mu","tau")))
R> for (i in 1:Nsim) {
+    # generate data:
+    mu <- rnorm(n=1, mean=mupriormean, sd=mupriorsd)  # effect
+    tau <- rhalfnormal(n=1, scale=taupriorscale)      # heterogeneity
+    k <- sample(c(2,3,5,10,20), size=1)               # number of studies
+    sigma <- runif(n=k, min=0.2, max=1.0)             # standard errors
+    y <- rnorm(n=k, mean=mu, sd=sqrt(sigma^2+tau^2))  # estimates
+    # perform analysis:
+    bma <- try(bayesmeta(y=y, sigma=sigma,
+      tau.prior=function(t){dhalfnormal(t, scale=taupriorscale)},
+      mu.prior=c(mupriormean, mupriorsd)))
+    # log probability integral transform (PIT) values:
+    if (!is.element("try-error", class(bma))) {
+      pit[i,"mu"]  <- bma$pposterior(mu=mu)
+      pit[i,"tau"] <- bma$pposterior(tau=tau)
+    }
+  }
```

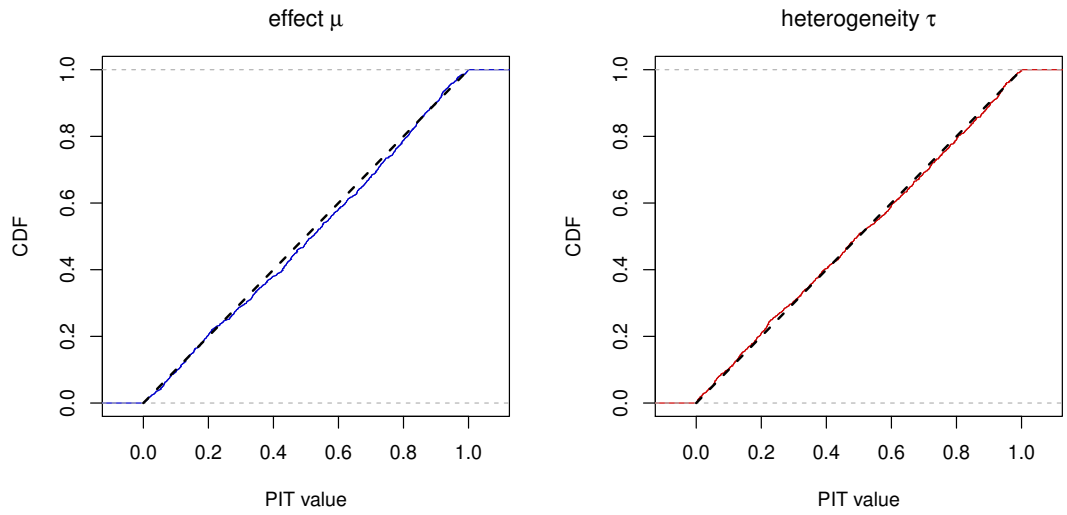

Figure 8: Empirical cumulative distribution functions of a sample of PIT values for effect and heterogeneity parameters. If the analysis is properly calibrated, the PIT values should follow a uniform distribution (black dashed line).

Prior parameters are set at the beginning, and matching settings are used for the analysis. Sample sizes (k) here are varied between 2 and 20, and standard errors (σ_i) between 0.2 and 1.0. The "`pit`" matrix consists of two column vectors of PIT values for the marginal effect (μ) and heterogeneity (τ) posteriors, respectively. We may now illustrate the empirical cumulative distribution function of the 1000 PIT values e.g. using the following commands:

```
R> plot(ecdf(pit[,"mu"]), col="blue",
+    main="effect (mu)", xlab="PIT value", ylab="CDF")
R> lines(0:1, 0:1, lty="dashed", lwd=2)
R> plot(ecdf(pit[,"tau"]), col="red",
+    main="heterogeneity (tau)", xlab="PIT value", ylab="CDF")
R> lines(0:1, 0:1, lty="dashed", lwd=2)
```

(see Fig. 8). The black dashed lines here indicate the limiting uniform distribution that should be approached for large numbers of simulations. The empirical distribution is in close agreement with the uniform distribution here.

What one can read off from the plots directly is the empirical coverage of one-sided upper credible limits. For example, one-sided 95% credible limits empirically exhibited a coverage of close to 95% in the simulations. For the heterogeneity, a curve above the main diagonal may be interpreted as "conservative" (in the sense of a tendency to overestimate heterogeneity), while for the effect, a conservative procedure should yield a curve below the diagonal at the lower end and above the diagonal at the upper end (i.e., leading to intervals that tend to be wider than necessary).

Acknowledgements

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement number FP HEALTH 2013-602144 through the *InSPiRe* project. Much of the present work evolved in close collaboration with Tim Friede, Beat Neuenschwander and Simon Wandel. Many thanks for helpful comments go to Thomas Asendorf, Burak Günhan, Markus Harden, Judith Heinz, Barbora Kessel, Tobias Mütze, Sibylle Sturtz, Steffen Unkel, and an anonymous reviewer.

References

Bartholomew DJ (1965). "A Comparison of Some Bayesian and Frequentist Inferences." *Biometrika*, **52**(1–2), 19–35. doi:10.2307/2333809.

Bayarri MJ, Berger JO (2004). "The Interplay of Bayesian and Frequentist Analysis." *Statistical Science*, **19**(1), 58–80. doi:10.1214/088342304000000116.

Bender R, Kuß O, Koch A, Schwenke C, Hauschke D (2014). "Application of Prediction Intervals in Meta-Analyses with Random Effects." Joint Statement of IQWiG, GMDS and IBS-DR. URL: https://www.iqwig.de/download/2014-03-07_Joint_Statement_Prediction_Intervals.pdf.

Berger J, Pericchi LR (2001). "Objective Bayesian Methods for Model Selection: Introduction and Comparison." In P Lahiri (ed.), *Model Selection*, Volume 38 of *IMS Lecture Notes*, pp. 135–193. Institute of Mathematical Statistics, Beachwood, OH. doi:10.1214/lnms/1215540968.

Berger JO (1985). *Statistical Decision Theory and Bayesian Analysis*. 2nd edition. Springer-Verlag, New York.

Berger JO, Deely J (1988). "A Bayesian Approach to Ranking and Selection of Related Means with Alternatives to Analysis-of-Variance Methodology." *Journal of the American Statistical Association*, **83**(402), 364–373. doi:10.1080/01621459.1988.10478606.

Berkhof J, van Mechelen I, Hoijting H (2000). "Posterior Predictive Checks: Principles and Discussion." *Computational Statistics*, **15**(3), 337–354. doi:10.1007/s001800000038.

Bodnar O, Link A, Arendacká B, Possolo A, Elster C (2017). "Bayesian Estimation in Random Effects Meta-Analysis Using a Non-Informative Prior." *Statistics in Medicine*, **36**(2), 378–399. doi:10.1002/sim.7156.

Bodnar O, Link A, Elster C (2016). "Objective Bayesian Inference for a Generalized Marginal Random Effects Model." *Bayesian Analysis*, **11**(1), 25–45. doi:10.1214/14-BA933.

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (2009). *Introduction to Meta-Analysis*. John Wiley & Sons.

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (2010). "A Basic Introduction to Fixed-Effect and Random-Effects Models for Meta-Analysis." *Research Synthesis Methods*, **1**(2), 97–111. doi:10.1002/jrsm.12.

Browne WJ, Draper D (2006). "A Comparison of Bayesian and Likelihood-Based Methods for Fitting Multilevel Models." *Bayesian Analysis*, **1**(3), 473–514. doi:10.1214/06-BA117.

Burr D (2012). "bspmma: An R package for Bayesian Semiparametric Models for Meta-Analysis." *Journal of Statistical Software*, **50**(4). doi:10.18637/jss.v050.i04.

Chalmers I, Hedges LV, Cooper H (2002). "A Brief History of Research Synthesis." *Evaluation & the Health Professions*, **25**(1), 12–37. doi:10.1177/0163278702025001003.

Chambert T, Rotella JJ, Higgs MD (2014). "Use of Posterior Predictive Checks as an Inferential Tool for Investigating Individual Heterogeneity in Animal Population Vital Rates." *Ecology and Evolution*, **4**(8), 1389–1397. doi:10.1002/ece3.993.

Chung Y, Rabe-Hesketh S, Choi IH (2013). "Avoiding Zero Between-Study Variance Estimates in Random-Effects Meta-Analysis." *Statistics in Medicine*, **32**(23), 4071–4089. doi:10.1002/sim.5821.

Cook SR, Gelman A, Rubin DB (2006). "Validation of Software for Bayesian Models Using Posterior Quantiles." *Journal of Computational and Graphical Statistics*, **15**(3), 675–692. doi:10.1198/106186006X136976.

Crins ND, Röver C, Goralczyk AD, Friede T (2014). "Interleukin-2 Receptor Antagonists for Pediatric Liver Transplant Recipients: A Systematic Review and Meta-Analysis of Controlled Studies." *Pediatric Transplantation*, **18**(8), 839–850. doi:10.1111/petr.12362.

Daniels M (1999). "A Prior for the Variance in Hierarchical Models." *The Canadian Journal of Statistics*, **27**(3), 567–578. doi:10.2307/3316112.

Davey J, Turner RM, Clarke MJ, Higgins JPT (2011). "Characteristics of Meta-Analyses and Their Component Studies in the Cochrane Database of Systematic Reviews: A Cross-Sectional, Descriptive Analysis." *BMC Medical Research Methodology*, **11**(1), 160. doi:10.1186/1471-2288-11-160.

Dawid AP (1982). "The Well-Calibrated Bayesian." *Journal of the American Statistical Association*, **77**(379), 605–610. doi:10.1080/01621459.1982.10477856.

Debray T, de Jong V (2012). *metamisc: Diagnostic and Prognostic Meta-Analysis*. R package, URL https://cran.r-project.org/package=metamisc.

Deeks JJ (2002). "Issues in the Selection of a Summary Statistic for Meta-Analysis of Clinical Trials with Binary Outcomes." *Statistics in Medicine*, **21**(11), 1575–1600. doi:10.1002/sim.1188.

Ding T, Baio G (2015). *bmeta: Bayesian Meta-Analysis and Meta-Regression*. R package, URL https://cran.r-project.org/package=bmeta.

DuMouchel WH, Normand SL (2000). "Computer Modeling Strategies for Meta-Analysis." In DK Stangl, DA Berry (eds.), *Meta-Analysis in Medicine and Health Policy*, pp. 127–178. CRC Press.

Follmann DA, Proschan MA (1999). "Valid Inference in Random Effects Meta-Analysis." *Biometrics*, **55**(3), 732–737. doi:10.1111/j.0006-341X.1999.00732.x.

Friede T, Röver C, Wandel S, Neuenschwander B (2017a). "Meta-Analysis of Few Small Studies in Orphan Diseases." *Research Synthesis Methods*, **8**(1), 79–91. doi:10.1002/jrsm.1217.

Friede T, Röver C, Wandel S, Neuenschwander B (2017b). "Meta-Analysis of Two Studies in the Presence of Heterogeneity with Applications in Rare Diseases." *Biometrical Journal*, **59**(4), 658–671. doi:10.1002/bimj.201500236.

Friedrich T, Knapp G (2013). "Generalized Interval Estimation in the Random Effects Meta Regression Model." *Computational Statistics & Data Analysis*, **64**, 165–179. doi:10.1016/j.csda.2013.03.011.

Ganschow R, Grabhorn E, Schulz A, von Hugo A, Rogiers X, Burdelski A (2005). "Long-Term Results of Basiliximab Induction Immunosuppression in Pediatric Liver Transplant Recipients." *Pediatric Transplantation*, **9**(6), 741–745. doi:10.1111/j.1399-3046.2005.00371.x.

Gelman A (2003). "A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-Fit Testing." *International Statistical Review*, **71**(2), 369–382. doi:10.1111/j.1751-5823.2003.tb00203.x.

Gelman A (2006). "Prior Distributions for Variance Parameters in Hierarchical Models." *Bayesian Analysis*, **1**(3), 515–534. doi:10.1214/06-BA117A.

Gelman A (2013). "Two Simple Examples for Understanding Posterior p-Values Whose Distributions are Far From Uniform." *Electronic Journal of Statistics*, **7**, 2595–2602. doi:10.1214/13-EJS854.

Gelman A, Carlin JB, Stern H, Dunson DB, Vehtari A, Rubin DB (2014). *Bayesian Data Analysis*. 3rd edition. Chapman & Hall / CRC, Boca Raton.

Gelman A, Meng XL, Stern H (1996). "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies." *Statistica Sinica*, **6**(4), 733–760. URL http://www.jstor.org/stable/24306036.

Gibelli NEM, Pinho-Apezzato ML, Miyatani HT, Maksoud-Filho JG, Silva MM, Ayoub AAR, Santos MM, Velhote MCP, Tannuri U, Maksoud JG (2004). "Basiliximab-Chimeric Anti-IL2-R Monoclonal Antibody in Pediatric Liver Transplantation: Comparative Study." *Transplantation Proceedings*, **36**(4), 956–957. doi:10.1016/j.transproceed.2004.04.070.

Gilks WR, Richardson S, Spiegelhalter DJ (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall / CRC, Boca Raton.

Gneiting T, Balabdaoui F, Raftery AE (2007). "Probabilistic Forecasts, Calibration and Sharpness." *Journal of the Royal Statistical Society B*, **69**(2), 243–268. doi:10.1111/j.1467-9868.2007.00587.x.

Gordon M, Lumley T (2017). *forestplot: Advanced Forest Plot Using 'Grid' Graphics*. R package version 1.7.2, URL http://cran.r-project.org/package=forestplot.

Gras JM, Gerkens S, Beguin C, Janssen M, Smets F, Otte JB, Sokal E, Reding R (2008). "Steroid-Free, Tacrolimus-Basiliximab Immunosuppression in Pediatric Liver Transplantation: Clinical and Pharmacoeconomic Study in 50 Children." *Liver Transplantation*, **14**(4), 469–477. doi:10.1002/lt.21397.

Gregory PC (2005). *Bayesian Logical Data Analysis for the Physical Sciences*. Cambridge University Press, Cambridge.

Guddat C, Grouven U, Bender R, Skipka G (2012). "A Note on the Graphical Presentation of Prediction Intervals in Random-Effects Meta-Analyses." *Systematic Reviews*, **1**(34). doi:10.1186/2046-4053-1-34.

Günhan BK (2017). *nmaINLA: Network Meta-Analysis Using Integrated Nested Laplace Approximations*. R package, URL http://cran.r-project.org/package=nmaINLA.

Guo J, Riebler A (2015). *meta4diag: Meta-Analysis for Diagnostic Test Studies*. R package, URL https://cran.r-project.org/package=meta4diag.

Hartung J, Knapp G (2001a). "On Tests of the Overall Treatment Effect in Meta-Analysis with Normally Distributed Responses." *Statistics in Medicine*, **20**(12), 1771–1782. doi:10.1002/sim.791.

Hartung J, Knapp G (2001b). "A Refined Method for the Meta-Analysis of Controlled Clinical Trials with Binary Outcome." *Statistics in Medicine*, **20**(24), 3875–3889. doi:10.1002/sim.1009.

Hartung J, Knapp G, Sinha BK (2008). *Statistical Meta-Analysis with Applications*. John Wiley & Sons, Hoboken, NJ, USA.

Heck DW, Gronau QF, Wagenmakers EJ (2017). *metaBMA: Bayesian Model Averaging for Random and Fixed Effects Meta-Analysis*. R package, URL https://cran.r-project.org/package=metaBMA.

Hedges LV, Olkin I (1985). *Statistical Methods for Meta-Analysis*. Academic Press, San Diego, CA, USA.

Heffron TG, Pillen T, Smallwood GA, Welch D, Oakley B, Romero R (2003). "Pediatric Liver Transplantation with Daclizumab Induction Therapy." *Transplantation*, **75**(12), 2040–2043. doi:10.1097/01.TP.0000065740.69296.DA.

Higgins JPT, Green S (eds.) (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.1.0, URL http://handbook.cochrane.org/.

Higgins JPT, Thompson SG (2002). "Quantifying Heterogeneity in a Meta-Analysis." *Statistics in Medicine*, **21**(11), 1539–1558. doi:10.1002/sim.1186.

Higgins JPT, Thompson SG, Spiegelhalter DJ (2009). "A Re-Evaluation of Random-Effects Meta-Analysis." *Journal of the Royal Statistical Society A*, **172**(1), 137–159. doi:10.1111/j.1467-985X.2008.00552.x.

Hoekstra R, Morey RD, Rouder JN, Wagenmakers EJ (2014). "Robust Misinterpretation of Confidence Intervals." *Psychonomic Bulletin & Review*, **21**(5), 1157–1164. doi:10.3758/s13423-013-0572-3.

Jaynes ET (1968). "Prior Probabilities." *IEEE Transactions on Systems Science and Cybernetics*, **SEC-4**(3), 227–241. doi:10.1109/TSSC.1968.300117.

Jaynes ET (1976). "Confidence Intervals vs. Bayesian Intervals." In WL Harper, CA Hooker (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, pp. 175–257. D. Reidel, Dordrecht. doi:10.1007/978-94-010-1436-6_6.

Jaynes ET (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge.

Jeffreys H (1946). "An Invariant Form for the Prior Probability in Estimation Problems." *Proceedings of the Royal Society of London, Series A*, **186**(1007), 453–461. doi:10.1098/rspa.1946.0056.

Jeffreys H (1961). *Theory of Probability*. 3rd edition. Clarendon Press, Oxford.

Johnson NL, Kotz S, Balakrishnan N (1994). *Continuous Univariate Distributions*. 2nd Edition. John Wiley & Sons, New York.

Kacker RN, Forbes A, Kessel R, Sommer KD (2008). "Bayesian Posterior Predictive p-Value of Statistical Consistency in Interlaboratory Evaluations." *Metrologia*, **45**(5), 512–523. doi:10.1088/0026-1394/45/5/004.

Kass RE, Wasserman L (1995). "A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion." *Journal of the American Statistical Association*, **90**(431), 928–934. doi:10.2307/2291327.

Kass RE, Wasserman L (1996). "The Selection of Prior Distributions by Formal Rules." *Journal of the American Statistical Association*, **91**(435), 1343–1370. doi:10.1080/01621459.1996.10477003.

Kontopantelis E, Springate DA, Reeves D (2013). "A Re-Analysis of the Cochrane Library Data: The Dangers of Unobserved Heterogeneity in Meta-Analyses." *PLoS ONE*, **8**(7), e69930. doi:10.1371/journal.pone.0069930.

Kruschke JK, Liddell TM (2018). "The Bayesian New Statistics: Hypothesis Testing, Estimation, Meta-Analysis, and Power Analysis From a Bayesian Perspective." *Psychonomic Bulletin & Review*, **25**(1), 178–206. doi:10.3758/s13423-016-1221-4.

Lewis S, Clarke M (2001). "Forest Plots: Trying to See the Wood and the Trees." *BMJ*, **322**(7300), 1479–1480. doi:10.1136/bmj.322.7300.1479.

Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, Clarke M, Devereaux PJ, Kleijnen J, Moher D (2009). "The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies that Evaluate Health Care Interventions: Explanation and Elaboration." *PLoS Medicine*, **6**(7), e1000100. doi:10.1371/journal.pmed.1000100.

Lindley DV (1977). "The Distinction Between Inference and Decision." *Synthese*, **36**(1), 51–58. doi:10.1007/BF00485691.

Lindsay BG (1995). *Mixture Models: Theory, Geometry and Applications*, Volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, Hayward, CA, USA. URL http://www.jstor.org/stable/4153184.

Lunn D, Barrett J, Sweeting A, Thompson S (2013). "Fully Bayesian Hierarchical Modelling in Two Stages With Application to Meta-Analysis." *Journal of the Royal Statistical Society C*, **62**(4), 551–572. doi:10.1111/rssc.12007.

Luo S, Chen Y, Su X, Chu H (2014). "mmeta: An R Package for Multivariate Meta-Analysis." *Journal of Statistical Software*, **56**(11). doi:10.18637/jss.v056.i11.

Mandelkern M (2002). "Setting Confidence Intervals for Bounded Parameters." *Statistical Science*, **17**(2), 149–172. doi:10.1214/ss/1030550859.

Meng XL (1994). "Posterior Predictive p-Values." *The Annals of Statistics*, **22**(3), 1142–1160. doi:10.1214/aos/1176325622.

Metropolis N, Ulam S (1949). "The Monte Carlo Method." *Journal of the American Statistical Association*, **44**(247), 335–341. doi:10.1080/01621459.1949.10483310.

Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers EJ (2016). "The Fallacy of Placing Confidence in Confidence Intervals." *Psychonomic Bulletin & Review*, **23**(1), 103–123. doi:10.3758/s13423-015-0947-8.

O'Hagan A, Pericchi L (2012). "Bayesian Heavy-Tailed Models and Conflict Resolution: A Review." *Brazilian Journal of Probability and Statistics*, **26**(4), 372–401. doi:10.1214/11-BJPS164.

Plummer M (2003). "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling." In K Hornik, F Leisch, A Zeileis (eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. Vienna, Austria.

Plummer M (2008). *rjags: Bayesian Graphical Models Using MCMC*. R package, URL http://cran.r-project.org/package=rjags.

Polson NG, Scott JG (2012). "On the Half-Cauchy Prior for a Global Scale Parameter." *Bayesian Analysis*, **7**(4), 887–902. doi:10.1214/12-BA730.

Poole C, Greenland S (1999). "Random-Effects Meta-Analyses are not Always Conservative." *American Journal of Epidemiology*, **150**(5), 469–475. doi:10.1093/oxfordjournals.aje.a010035.

Pullenayegum EM (2011). "An Informed Reference Prior for Between-Study Heterogeneity in Meta-Analyses of Binary Outcomes." *Statistics in Medicine*, **30**(26), 3082–3094. doi:10.1002/sim.4326.

Rhodes KM, Turner RM, Higgins JPT (2015). "Predictive Distributions were Developed for the Extent of Heterogeneity in Meta-Analyses of Continuous Outcome Data." *Journal of Clinical Epidemiology*, **68**(1), 52–60. doi:10.1016/j.jclinepi.2014.08.012.

Riley RD, Higgins JP, Deeks JJ (2011). "Interpretation of Random Effects Meta-Analyses." *BMJ*, **342**, d549. doi:10.1136/bmj.d549.

Röver C (2015). "bayesmeta: Bayesian Random-Effects Meta Analysis." R package. URL: http://cran.r-project.org/package=bayesmeta.

Röver C, Friede T (2017). "Discrete Approximation of a Mixture Distribution via Restricted Divergence." *Journal of Computational and Graphical Statistics*, **26**(1), 217–222. doi:10.1080/10618600.2016.1276840.

Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B (2014). "Robust Meta-Analytic-Predictive Priors in Clinical Trials with Historical Control Information." *Biometrics*, **70**(4), 1023–1032. doi:10.1111/biom.12242.

Schmidli H, Neuenschwander B, Friede T (2017). "Meta-Analytic-Predictive use of Historical Variance Data for the Design and Analysis of Clinical Trials." *Computational Statistics & Data Analysis*, **113**, 100–110. doi:10.1016/j.csda.2016.08.007.

Schuller S, Wiederkehr JC, Coelho-Lemos IM, Avilla SG, Schultz C (2005). "Daclizumab Induction Therapy Associated with Tacrolimus-MMF has Better Outcome Compared with Tacrolimus-MMF Alone in Pediatric Living Donor Liver Transplantation." *Transplantation Proceedings*, **37**(2), 1151–1152. doi:10.1016/j.transproceed.2005.01.023.

Schwarzer G (2007). "meta: An R Package for Meta-Analysis." *R News*, **7**(3), 40–45.

Schwarzer G, Carpenter JR, Rücker G (2015). *Meta-Analysis with R*. Springer-Verlag.

Seidel WE (2010). "Mixture Models." In M Lovric (ed.), *International Encyclopedia of Statistical Science*, pp. 827–829. Springer-Verlag, Heidelberg. doi:10.1007/978-3-642-04898-2.

Senn S (2007). "Trying to be Precise About Vagueness." *Statistics in Medicine*, **26**(7), 1417–1430. doi:10.1002/sim.2639.

Severini TA (1991). "On the Relationship Between Bayesian and Non-Bayesian Interval Estimation." *Journal of the Royal Statistical Society B*, **53**(3), 611–618. URL http://www.jstor.org/stable/2345590.

Shaked M, Shanthikumar JG (2007). *Stochastic Orders*. Springer-Verlag, New York.

Sidik K, Jonkman JN (2002). "A Simple Confidence Interval for Meta-Analysis." *Statistics in Medicine*, **21**(21), 3153–3159. doi:10.1002/sim.1262.

Sinharay S, Johnson MJ, Stern HS (2006). "Posterior Predictive Assessment of Item Response Theory Models." *Applied Psychological Measurement*, **30**(4), 298–321. doi:10.1177/0146621605285517.

Smith TC, Spiegelhalter DJ, Thomas A (1995). "Bayesian Approaches to Random-Effects Meta-Analysis: A Comparative Study." *Statistics in Medicine*, **14**(24), 2685–2699. doi:10.1002/sim.4780142408.

Spada M, Petz W, Bertani A, Riva S, Sonzogni A, Giovannelli M, Torri E, Torre G, Colledan M, Gridelli B (2006). "Randomized Trial of Basiliximab Induction Versus Steroid Therapy in Pediatric Liver Allograft Recipients Under Tacrolimus Immunosuppression." *American Journal of Transplantation*, **6**(8), 1913–1921. doi:10.1111/j.1600-6143.2006.01406.x.

Spence GT, Steinsaltz D, Fanshawe TR (2016). "A Bayesian Approach to Sequential Meta-Analysis." *Statistics in Medicine*, **35**(29), 5356–5375. doi:10.1002/sim.7052.

Spiegelhalter DJ (2004). "Incorporating Bayesian Ideas Into Health-Care Evaluation." *Statistical Science*, **19**(1), 156–174. doi:10.1214/088342304000000080.

Spiegelhalter DJ, Abrams KR, Myles JP (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. John Wiley & Sons.

Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR (1999). "An Introduction to Bayesian Methods in Health Technology Assessment." *BMJ*, **319**, 508–512. doi:10.1136/bmj.319.7208.508.

Sutton AJ, Abrams KR (2001). "Bayesian Methods in Meta-Analysis and Evidence Synthesis." *Statistical Methods in Medical Research*, **10**(4), 277–303. doi:10.1177/096228020101000404.

Sweeting MJ, Sutton AJ, Lambert PC (2004). "What to Add to Nothing? Use and Avoidance of Continuity Corrections in Meta-Analysis of Sparse Data." *Statistics in Medicine*, **23**(9), 1351–1375. doi:10.1002/sim.1761.

Szucs D, Ioannidis JPA (2017). "When Null Hypothesis Significance Testing is Unsuitable for Research: A Reassessment." *Frontiers in Human Neuroscience*, **11**, 390. doi:10.3389/fnhum.2017.00390.

The Cochrane Collaboration (2014). *Review Manager (RevMan)*. Copenhagen. Version 5.3, URL http://ims.cochrane.org/revman.

Tibshirani R (1989). "Noninformative Priors for One Parameter of Many." *Biometrika*, **76**(3), 604–608. doi:10.1093/biomet/76.3.604.

Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JPT (2012). "Predicting the Extent of Heterogeneity in Meta-Analysis, Using Empirical Data from the Cochrane Database of Systematic Reviews." *International Journal of Epidemiology*, **41**(3), 818–827. doi:10.1093/ije/dys041.

Turner RM, Jackson D, Wei Y, Thompson SG, Higgins PT (2015). "Predictive Distributions for Between-Study Heterogeneity and Simple Methods for Their Application in Bayesian Meta-Analysis." *Statistics in Medicine*, **34**(6), 984–998. doi:10.1002/sim.6381.

van de Schoot R, Kaplan D, Denissen J, Asendorpf JB, Neyer FJ, van Aken MAG (2014). "A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research." *Child Development*, **85**(3), 842–860. doi:10.1111/cdev.12169.

van den Noortgate W, Onghena P (2005). "Parametric and Nonparametric Bootstrap Methods for Meta-Analysis." *Behavior Research Methods*, **37**(1), 11–22. doi:10.3758/BF03206394.

van Erp S, Verhagen J, Grasman RPPP, Wagenmakers EJ (2017). "Estimates of Between-Study Heterogeneity for 705 Meta-Analyses Reported in *Psychological Bulletin* from 1990–2013." *Journal of Open Psychology Data*, **5**(4). doi:10.5334/jopd.33.

Verde PE (2011). *bamdit: Bayesian Meta-Analysis of Diagnostic Test Data*. R package, URL https://cran.r-project.org/package=bamdit.

Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, Kuß O, Higgins JPT, Langan D, Salanti G (2016). "Methods to Estimate the Between-Study Variance and its Uncertainty in Meta-Analysis." *Research Synthesis Methods*, **7**(1), 55–79. doi:10.1002/jrsm.1164.

Viechtbauer W (2010). "Conducting Meta-Analyses in R with the metafor Package." *Journal of Statistical Software*, **36**(3). doi:10.18637/jss.v036.i03.

Wandel S, Neuenschwander B, Röver C, Friede T (2017). "Using Phase II Data for the Analysis of Phase III Studies: an Application in Rare Diseases." *Clinical Trials*, **14**(3), 277–285. doi:10.1177/1740774517699409.

Wasserstein RL (2016). "ASA Statement on Statistical Significance and *p*-Values." *The American Statistician*, **70**(2), 131–133. doi:10.1080/00031305.2016.1154108.

Affiliation:

Christian Röver
Department of Medical Statistics
University Medical Center Göttingen
Georg-August-Universität
37073 Göttingen, Germany
E-mail: christian.roever@med.uni-goettingen.de
URL: http://www.gwdg.de/~croever
ORCID iD: 0000-0002-6911-698X

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* 2017-11-23
doi:10.18637/jss.v000.i00 *Accepted:* 2018-09-02