A bias-corrected meta-analysis model for combining, studies of different types and quality

Pablo Emilio Verde

Abstract
Public health researchers may have to decide whether to perform a meta-analysis including only high-quality randomized clinical trials (RCTs) or whether to include a mixture of all the available evidence, namely RCTs of varying quality and observational studies (OS). The main hurdle when combining disparate evidence in a meta-analysis is that we are not only combining results of interest but we are also combining multiple biases. Therefore, commonly applied meta-analysis methods may lead to misleading conclusions.

In this paper, we present a new Bayesian hierarchical model, called the bias-corrected (BC) meta-analysis model, to combine different study types in meta-analysis. This model is based on a mixture of two random effects distributions, where the first component corresponds to the model of interest and the second component to the hidden bias structure. In this way, the resulting model of interest is adjusted by the internal validity bias of the studies included in a systematic review.

We illustrate the BC model with two meta-analyses: The first one combines RCTs and OS to assess effectiveness of vaccination to prevent invasive pneumococcal disease. The second one investigates the effectiveness of stem cell treatment in heart disease patients. Our results show that ignoring internal validity bias in a meta-analysis may lead to misleading conclusions. However, if a meta-analysis model contemplates a bias adjustment, then RCTs results may increase their precision by including OS in the analysis. The BC model has been implemented in JAGS and R, which facilitate its application in practice.

KEYWORDS
Bayesian hierarchical models, comparative effectiveness methods, conflict of evidence, generalized evidence synthesis, meta-analysis

1 INTRODUCTION

Health researchers may have to decide whether to perform a meta-analysis including only results of randomized clinical trials (RCTs) or a meta-analysis of all available evidence, which may include a mixture of RCTs of varying quality and
TABLE 1 Results of efficacy and effectiveness of PPV23 vaccination on elderly patients against IPD: three RCT (randomized control trial) studies; three case-control studies, and five cohort studies. Effectiveness is assessed by the log(OR), where values of log(OR) lower than zero favor vaccination. The cohort study, Tsai 2015, presented a perfect balance between treatment groups, which had been induced by propensity score matching. The RCT study, Honkanen 1999, excluded the endpoint pneumococcal pneumonia, which resulted in “unclear” risk of bias evaluation.

| Study            | Design     | log(OR) | SE   | Total number of vaccine | Total PPV23 | Risk of bias |
|------------------|------------|---------|------|-------------------------|-------------|--------------|
| Honkanen 1999    | RCT        | −0.97   | 0.84 | 18,488                  | 19,549      | Unclear      |
| Maruyama 2010    | RCT        | −1.94   | 1.51 | 1,149                   | 1,140       | Low          |
| Ortqvist 1998    | RCT        | −1.52   | 1.10 | 873                     | 793         | Low          |
| Hechter 2012     | Cohort     | −1.04   | 0.89 | 27,320                  | 3,962       | High         |
| Jackson 2003     | Cohort     | −0.58   | 0.26 | 42,977                  | 84,203      | Low          |
| Ochoa-Gondar 2014| Cohort     | −0.96   | 0.75 | 30,000                  | 28,662      | Low          |
| Tsai 2015        | Cohort     | −1.42   | 0.33 | 229,181                 | 229,181     | High         |
| Vila-Corcoles 2006| Cohort | −0.51   | 0.51 | 16,504                  | 17,401      | Low          |
| Dominguez 2005   | Case-control | −1.18  | 0.26 | 447                     | 149         | Low          |
| Leventer-Roberts 2015 | Case-control | −0.54 | 0.17 | 848                     | 212         | Low          |
| Vila-Corcoles 2009 | Case-control | −1.08  | 0.34 | 188                     | 94          | Low          |

observational studies (OS). While the evidence of high-quality RCTs is the gold standard to prove clinical efficacy, it may happen that for a particular treatment this evidence may not exist. Moreover, it may happen that the settings of RCTs experimentation (e.g., inclusion/exclusion criteria of participants) do not reflect the context of real-world applications. Therefore, by combining different study types we may improve our understanding of the evidence and scope of a meta-analysis result. Table 1 illustrates this situation in a recent meta-analysis (Falkenhorst et al., 2017), where RCTs, cohort, and case-control studies were included to infer effectiveness of vaccination against invasive pneumococcal disease (IPD) in elderly patients. Effectiveness is assessed by the odds ratio in the logarithmic scale (log(OR)), where values lower than zero favor vaccination.

A first glance at Table 1 shows that all studies reported negative values of log(OR), with larger standard errors (SE) in the RCTs’ results and a substantial imbalance between the number of participants in each treatment group in the cohorts and case-control studies. These differences between study types may raise the question of potential biases and conflict of evidence between the studies. Therefore, combining studies’ results and ignoring study types may lead to misleading conclusions. However, like in this example, analyzing results separately by study types may discard potential common aspects between studies, including the assessment of effectiveness on the same intervention, by using the same outcome on similar patient populations.

How can we correct the potential bias effect of the OS to improve the evidence given by the RCTs? The solid line in Figure 1 presents a promising solution, which corresponds to the posterior distribution of the pooled log(OR) that has been corrected by internal validity bias, that is to say the studies’ characteristics that potentially affect the quality of the design and conduct of the studies.

The model used to calculate this posterior distribution, called the BC meta-analysis model, is the topic of this paper and is presented in Section 2. The BC model is built upon two submodels: a model of interest and a model of bias. The main idea is that by jointly incorporating these two models, it is possible to isolate the model of interest, which has been corrected by internal validity bias. In Figure 1, this approach produces a posterior distribution that is centered close to the posterior mean of a meta-analysis using only RCTs’ data and that has borrowed information from the OS results.

One important advantage of this approach is that the proportion of OS studies (or studies at high risk of bias) is used as a simple prior elicitation of the probability of bias in the meta-analysis. Furthermore, in Section 2, we extend the model to include study characteristics that could explain the propensity of bias of each study.

A review of methods and applications of combining randomized and observational evidence over the last two decades is presented by Verde and Ohmann (2015), where we classified statistical approaches in four main groups: The confidence profile method (CPM) (Eddy, Hasselblad, & Shachter, 1992), cross-design synthesis (CDS) (Droitcour, Silberman, & Chelinsky, 1993), direct likelihood bias modeling, and Bayesian hierarchical modeling.
The CPM was introduced by Eddy et al. (1992) as a general statistical framework to combine multiple sources of information in evidence synthesis. Although the CPM is not currently applied in health research, it has influenced methodological work during the last decades, including network meta-analysis, the use of Bayesian networks, and graphical models in evidence synthesis. The “real-world evidence synthesis” has produced “a revival” of the CDS’ ideas during the last years.

CDS (Droitcour et al., 1993) was a heuristic procedure to generalize RCTs’ external validity bias by using OS results. Kaizar (2011) proposed a simple CDS effect estimator within a formal framework of causal estimation. Recent applications of the CDS approach include the work of Greenhouse et al. (2017) and Vaitsiakhovich, Filonenko, Lynen, Endrikat, and Gerlinger (2018). Varadhan, Henderson, and Weiss (2016) and Henderson, Varadhan, and Weiss (2017) presented another approach for CDS that the authors called the calibrated risk-adjusted modeling (CRAM). The key feature of the CRAM is the identification of the heterogeneity in treatment effect, which is used to compute a calibration factor for unmeasured confounding for the OS relative to the RCT. The idea of the “direct likelihood modeling of bias” is to adjust studies’ internal and external validity bias by modifying the likelihood contribution of each study to the meta-analysis. In this way, the pooled treatment effect of the bias-adjusted meta-analysis reflects a more realistic estimate than the naive meta-analysis. However, this bias adjustment can be difficult to implement in practice. Turner, Spiegelhalter, Smith, and Thompson (2009) proposed that a group of assessors evaluate each study included in the systematic review and estimate the internal and external validity bias by using a score system. The pooled estimate of the bias components is used to adjust the likelihood contribution of each study. Schnell-Inderst et al. (2017) presented a case study combining results from RCTs and OS by using Turner’s procedure. Another direct likelihood modeling of bias is the application of a “power prior” that discounts for biased information in the combination of RCTs and OS (Ibrahim, Chen, Gwon, & Chen, 2015). For example, Welton, Sutton, Cooper, Abrams, and Ades (2012) built a power prior by discounting observational evidence and applied this prior to the relative treatment effect in a random effects meta-analysis of RCTs. The work of Ferguson, Alvarez-Iglesias, Newell, Hinde, and O’Donnell (2019) is another way to discount the likelihood contributions, where the authors proposed a weighted estimator of treatment effect as a convex combination between the estimates based on the RCTs and OS.

Bayesian hierarchical modeling has been the main stream of innovations in recent research of combining RCTs and OS. Verde, Ohmann, Morbach, and Icks (2016) and Verde (2019) presented a hierarchical meta-regression approach, a Bayesian approach that may help to generalize RCTs’ results to populations of patients not included in RCTs. Wang and Rosner (2019) investigated causality inference in real-world data by combining RCTs with prospectively designed
registry studies. The authors developed a novel Bayesian nonparametric approach for propensity scores. Efthimiou et al. (2017) investigated different Bayesian hierarchical models for combining randomized and nonrandomized evidence in network meta-analysis.

Disparate evidence in meta-analysis could also come from the diversity in the quality of RCTs. The method presented in this paper aims to handle this situation as well. There are two main approaches to combine RCTs with potential risk of bias in meta-analysis: the work of Welton, Ades, Carlin, Altman, and Sterne (2009) and of Dias et al. (2010). These approaches differ in the way that the bias component is estimated and incorporated into the meta-analysis model. Welton et al. (2009) presented a mixed-effects meta-analysis model to combine RCTs with different risk of bias. In this model, the fixed effect corresponds to the relative treatment effect and the random component to the bias effect. The authors started by assuming that each study could be evaluated as having a “high” or a “low” risk of bias. In this approach, the hyperparameters of the bias component are estimated by using empirical results coming from meta-epidemiological data consisting of a large number of meta-analyses. This approach assumes exchangeability between the meta-epidemiological data and the current meta-analysis. The idea of collecting objective prior information from meta-epidemiological data is attractive, but these meta-epidemiological studies are short in supply for practical work.

Dias et al. (2010) introduced a Bayesian hierarchical model to combine RCTs at different levels of risk of bias in a network meta-analysis. This model has the advantage of using only the information at hand without incorporating external information coming from meta-epidemiological studies. The authors presented several models to accommodate different situations of treatment comparisons in the network meta-analysis, for example, placebo vs. active treatment, active vs. active treatment, and so on. The authors performed a risk of bias assessment of each study, where a study's risk was classified as “high,” “low,” or “unclear.” The model of bias is a meta-regression random-effects component, where the resulting bias assessment is used as a binary covariate for the high/low classification. When a study was assessed as “unclear,” the authors handled this case as a missing class. The class imputation was performed by using the WinBUGS 1.4.3 software (Spiegelhalter, Thomas, Best, & Lunn, 2007), and in this case the covariate imputation model assumed a missing at random mechanism (Little & Rubin, 2019). Therefore, the predicted class was proportional to the majority class. What the authors called “the probability of bias” was actually “the probability of the missing class.” Like the approach of Dias et al. (2010), the model presented in this article only incorporates the information at hand to interpret and adjust for internal validity bias. However, these approaches are methodologically different.

We organized this paper as follows: In Section 2, we present methodological details of our approach. In Section 3, we present the application in real data examples with meta-analysis of different study types and RCTs with a diversity of quality. We conclude the paper in Section 4 by highlighting advantages and limitations of the model and by pointing out further methodological research in our approach.

## 2 THE BAYESIAN BIAS-CORRECTED META-ANALYSIS MODEL

In this section, we present a Bayesian model that we called the BC meta-analysis model. The BC model is based upon a mixture of two random effects distributions that represent a model of interest and a model of bias. Our hope is that by embedding the meta-analysis in a hierarchical Bayesian mixture model, it is possible to isolate the model of interest, which has been adjusted by internal validity bias.

Suppose that a meta-analysis of \( N \) studies has reported effect size estimates \( y_1, \ldots, y_N \) with their respective standard errors \( SE_1, \ldots, SE_N \). In addition, we assume that the sample size within studies is large enough to ignore uncertainty on \( SE_i \) (\( i = 1, 2, \ldots, N \)). We model \( y_i \) with a normal likelihood:

\[
y_i | \theta^B_i \sim N(\theta^B_i, SE_i^2),
\]

where \( \theta^B_i \) is the unobserved biased effect of study \( i \). We describe \( \theta^B_i \) as the combination of three hidden variables \( \theta_i, I_i, \) and \( \beta_i \):

\[
\theta^B_i = (1 - I_i) \theta_i + I_i (\theta_i + \beta_i).
\]

The indicator variable \( I_i \) expresses our uncertainty that study \( i \) is biased:

\[
I_i = \begin{cases} 
1 & \text{if study } i \text{ is biased} \\
0 & \text{otherwise,}
\end{cases}
\]
where the probability of internal validity bias is

\[ P(I_i = 1) = 1 - P(I_i = 1) = \pi_{\text{bias}}. \]

The parameter \( \pi_{\text{bias}} \) drives the assumption about bias in this model, that is if two mixture components are required in a meta-analysis. We start by modeling \( \pi_{\text{bias}} \) as

\[ \pi_{\text{bias}} \sim \text{Beta}(a_0, a_1), \quad (4) \]

where \( a_0 \) and \( a_1 \) indicate our prior information regarding internal validity bias of the studies. In Section 2.1, we give ways to elicit \( a_0 \) and \( a_1 \) and to perform a sensitivity analysis on the meta-analysis results. In Section 2.3, we extend the model to assess if a study’s characteristics may explain the propensity of bias.

The study effect \( \theta_i \) is the unobserved effect across unbiased studies and \( \beta_i \) the unobserved internal validity bias of study \( i \). We model \( \theta_i \) and \( \beta_i \) as independent effects with distributions:

\[ \theta_i \sim N(\mu, \tau^2), \quad \text{and} \quad \beta_i \sim N(\mu_{\beta_i}, \tau^2_{\beta_i}). \quad (5) \]

The distribution of \( \theta_i \) is indeed the model of interest. The parameter \( \mu \) corresponds to the mean effect across unbiased effects, and \( \tau \) represents the meta-analysis’ variability due to diversity in populations, interventions, exposures, etc.

The distribution of \( \beta_i \) represents our uncertainty regarding the internal validity bias of study \( i \). The parameters \( \mu_{\beta_i} \) and \( \tau^2_{\beta_i} \) account for the mean bias and variability in the quality of design and factors threatening the study’s integrity (patient selection bias, reporting bias, etc).

In order to make the model parameters identifiable, we make the assumption that the distribution of \( \beta_i \) across biased studies has an identical bias mean, that is \( \mu_{\beta_i} = \mu_{\beta} \). In addition, we assume that in a particular context of application, the internal validity bias may favor one treatment. For example, by evaluating the study types or other measures of internal validity we can suspect the direction of bias. Therefore, we assume that

\[ \mu_{\text{biased}} = \mu + \mu_{\beta}, \quad \text{with} \quad \mu_{\beta} > 0. \quad (6) \]

Equation (6) ensures that \( \mu \) and \( \mu_{\text{biased}} \) are identifiable parameters in this model.

Of course, in some situations, the direction of bias may be difficult to assess using the evidence reported by the studies. In Section 2.2, we explain a diagnostic plot to understand if bias’ direction strongly depends on prior assumptions or not.

We can combine the dispersion parameters \( \tau^2 \) and \( \tau^2_{\beta_i} \) as following:

\[ q_i = \frac{\tau^2}{\tau^2 + \tau^2_{\beta_i}}, \quad \text{and} \quad \beta_i \sim N(\mu_{\beta}, \tau^2_{\beta_i}). \quad (7) \]

the resulting random effects distribution of \( \theta_i^\beta \) is

\[ \theta_i^\beta | q_i \sim (1 - \pi_{\text{bias}}) N(\mu, \tau^2) + \pi_{\text{bias}} N(\mu_{\text{biased}}, \tau^2 q_i). \quad (8) \]

The quality weights \( q_i \) can be interpreted as the proportion of the between-studies’ variability that is not explained by internal validity bias. In practice, these weights are unknown, Spiegelhalter and Best (2003) proposed to give a set of fixed values to \( q_i \) (e.g., 0.1, 0.2, 0.3) to discount low-quality studies and to perform a sensitivity analysis. In this paper, we consider \( q_i \) as scale random variables with \( 0 < q_i < 1 \) and we model \( q_i \) as

\[ q_i \sim \text{Beta}(\nu, 1). \quad (9) \]

The marginal density of the bias component corresponds to a slash distribution with a location parameter \( \mu_{\text{biased}} \), a scale parameter \( \tau^2 \), and a shape parameter \( \nu \). This distribution has heavier tails than the normal distribution, where the parameter \( \nu \) controls the shape of the tails (Rogers & Tukey, 1972). The slash distribution has been used in adaptive robust
regression (Lange & Sinsheimer, 1993) and in Bayesian robust linear mixed models (Rosa, Padovani, & Gianola, 2003). In our application, the idea is to penalize the influence of studies which fall in the bias component.

The distribution (8) explicitly accounts for diversity and bias in the random effects, where we can extract the model of interest of $\hat{\theta}_i$. We are going to see in Section 3 that by combining and correcting for bias we can add important information to a set of good quality studies.

Figure 2 gives a schematic representation of the BC model by using a directed acyclic graph (DAG). The model of bias and the model of interest are presented on the left- and the right-hand side, respectively. We follow the usual graphical description, where ellipses represent random variables and double-lined ellipses indicate hyperprior parameters. A double-lined arrow indicates a functional relationship between variables and a single-lined arrow a stochastic relationship.

In Figure 2, we can see that the model of interest includes the BC pooled mean $\mu$, the BC study effect $\theta_i$, and the between studies variability $\tau$. Although the conclusions of a meta-analysis depend on the posterior distributions of $\mu$ and $\tau$, the model allows us to understand how the bias correction operates at the study effect $\theta_i$. We are going to illustrate this aspect of the model in Section 3.2.

### 2.1 Priors for hyperparameters

The model presented above is completed by the following independent weakly informative hyperparameters: $\mu$, $\tau$, $\mu_{\beta}$, $\pi_{\text{bias}}$, and $\nu$.

For the location parameters $\mu$ and $\mu_{\beta}$, we use $\mu \sim N(0, 100)$ and $\mu_{\beta} \sim \text{Uniform}(0, 10)$, respectively. These priors ensure the identifiability of $\mu$ and $\mu_{\beta}$.

For the scale parameter $\tau$, we use a half-Cauchy distribution with scale parameter $S$, that is, $\tau \sim HC(S)$. This density is flat for very small values ($\tau \ll S$) and has a heavy-tailed for large values ($\tau \gg S$). In this way, a misspecification of $S$ is...
harmless as it simply makes the prior flatter, whereas if $S$ is too small then the heavy tail ensures that enough data will dominate the prior. In Section 3, we use $S = 0.5$, this value gives a prior probability $P(\tau > S) = 0.5$ and $P(\tau > 4 \times S) = 0.16$. Therefore, this prior distribution covers possible values of $\tau$ in the log(OR) and mean difference scale.

For the probability of bias, we use $\pi_{\text{bias}} \sim \text{Beta}(a_0, a_1)$. We give two approaches to elicit $a_0$ and $a_1$: The first one is just a noninformative uniform prior, which results by taking $a_0 = a_1 = 1$. This prior corresponds to a prior ignorance regarding bias. We use this prior in the example presented in Section 3.2.

The second approach is informative. In this approach, we elicit the values of $a_0$ and $a_1$ by incorporating the number of different study types in the meta-analysis. The idea is to choose $a_0$ and $a_1$ in a way that the probability of bias is bounded by the proportion of OS. Let $N_{\text{OS}}$ be the number of OS in a meta-analysis, we choose $a_0$ and $a_1$ by taking the prior 90% quantile located at $N_{\text{OS}}/N$ and the prior median at $(N_{\text{OS}} - K)/N$. In other words, we concentrate the prior distribution at values lower than the proportion of OS studies. In Section 3.1, we take $K = 1$ which means that we give a 50% chance to include at least one OS in the BC group. In practice, we can perform a prior-to-posterior sensitivity analysis for different values of $K$. The same approach can be used if $N_{\text{ROB}}$ is the number of studies at high risk of bias.

To elicit a value of $\nu$, we can use the prior mean $q_i$, which is $E(q_i) = \nu / (\nu + 1)$.

In Section 3, we take $\nu = 0.5$, which corresponds to down-weighting in average $E(q_i) = 0.33$ low-quality studies. Alternately, for a meta-analysis with a large number of studies, a Gamma$(b_0, b_1)$ distribution with $b_0 < b_1$ can be adopted as a prior distribution for $\nu$, which has the advantage of being conditionally conjugated with (9) (Rosa et al., 2003).

2.2 Some model diagnostics

When we apply the BC model, it is important to understand if there is a point at which the $\mu_{\beta}$ or $\pi_{\text{bias}}$ are so low that no adjustment should be conducted. That is to say, we cannot isolate the model of interest from a two-component mixture model. This issue is related with deciding on the number of components in a mixture distribution, which is an important and difficult problem that has not been completely resolved yet (McLachlan & Peel, 2000).

To tackle this issue, we are going to display the jointly posterior distribution of the mean bias $\mu_{\beta}$ and the probability of bias $\pi_{\text{bias}}$ as a contour plot at different levels (50%, 75%, and 95%). In this way, we can assess the relationship between $\mu_{\beta}$ and $\pi_{\text{bias}}$. In addition, we display the marginal distributions of $\mu_{\beta}$ and $\pi_{\text{bias}}$ in the same plot. Another useful diagnostic quantity is the posterior probability of having a unique mode in the model. In a mixture of two normal distributions with different means $\mu_1$ and $\mu_2$ and different variances $\sigma_1^2$ and $\sigma_2^2$, the sufficient condition of having one mode is given by $|\mu_1 - \mu_2| < 2 \min(\sigma_1, \sigma_2)$ (McLachlan & Peel, 2000).

In the BC model, we are interested in the existence of a bias-correction component. Therefore, by taking $|\mu_{\beta}| = |\mu_1 - \mu_2|$ and given that $1 < q_i < 0$ the min$(\sigma_1, \sigma_2)$ is $\tau$, we quantify the existence of a bias component by the posterior probability that $Pr(|\mu_{\beta}| > 2 \tau |\text{Data})$. In addition, in the contour plot of the join posterior distribution of $\mu_{\beta}$ and $\pi_{\text{bias}}$, we display a reference line parallel to the $\pi_{\text{bias}}$ axes at the posterior mean $E(2 \tau |\text{Data})$.

2.3 Systematic change of the probability of bias

In the previous section, we assumed that each study has the same prior probability of bias $P(I_i = 1) = \pi_{\text{bias}}$. In Section 2.1, we showed how to elicit a prior distribution of $\pi_{\text{bias}}$ to incorporate a prior information into the model. However, it is more natural to assume that $P(I_i = 1)$ may vary from study to study and this variation could be systematically explained. In this section, we are going to formalize this idea.

Let $x_i$ be an outcome variable that measures internal validity of study $i$ ($i = 1, 2, ..., N$). For example, in Section 3.2, $x_i$ is the number of detected discrepancies in published trials. Then, we model the probability of bias conditioned to $x_i$ as follows:

$$
\log \left( \frac{P(I_i = 1|x_i)}{P(I_i = 0|x_i)} \right) = \alpha_0 + \alpha_1 x_i,
$$

(10)
where \( \exp(\alpha_0) \) represents the odds ratio of \( \pi_{\text{bias}} \) and \( \exp(\alpha_1) \) the odds of bias by units of \( x_1 \). When \( \alpha_1 > 0 \) indicates that an increase in \( x_1 \) is associated with an increase in the probability of bias. Clearly, other link functions can be used in Equation (10) (e.g., probit and cloglog). In the Supporting Information, we provide the implementation of different link functions for this model.

The main issue when modeling hidden variables is the difficulty of learning the posterior distributions of \( \alpha_0 \) and \( \alpha_1 \) from the data. Therefore, in a meta-analysis with a small number of studies, we can expect that the posterior distributions will be similar to the priors. However, we are going to see in Section 3.2 that in a meta-analysis with \( N = 31 \) studies, it is sensible to search for a systematic explanation of bias.

For this kind of analysis, we use weakly informative priors for the regression coefficients: \( \alpha_0 \sim N(0,10) \) and \( \alpha_1 \sim N(0,10) \). We propose to display the posterior distribution of (10) for a continuous variable \( x_1 \), by calculating (10) as a functional parameter for a grid of values of \( x \) and presenting their posterior intervals at each value of \( x \).

2.4 Modeling remarks

2.4.1 Contribution of the evidence at risk of bias

One interesting question rising from combining RCTs and OS is the amount of additional evidence gained by combining disparate data. One possible answer given by \( N^+ \), which is the effective number of studies falling into the unbiased class:

\[
N^+ = E(N(1 - \pi_{\text{bias}})|\text{Data}).
\] (11)

The posterior mean (11) can be compared with objective information, for example, the total number of RCTs in a meta-analysis of studies with different types.

2.4.2 Uncertainty in the direction of bias

It may happen that in a particular situation there is no clear evidence on the direction of bias. In this case, a simple strategy is to assume that \( \mu_\beta = 0 \), which results in a contaminated distribution of \( \theta_i^\beta \):

\[
\theta_i^\beta \sim (1 - \pi_{\text{bias}})N(\mu,\tau^2) + \pi_{\text{bias}} s(\mu,\tau,\nu),
\] (12)

where \( s(\mu,\tau,\nu) \) is a slash distribution with location \( \mu \), a scale \( \tau \), and a shape \( \nu \).

An alternative approach is to give the same priors to \( \mu \) and \( \mu_\beta \) and analyze if we have enough data to identify the direction of \( \mu_\beta \). The diagnostic techniques presented in Section 2.2 can be useful in this situation as well.

2.4.3 Using exact likelihood contributions

The use of a normal likelihood (1) has its perils (Jackson & White, 2018), in particular when the number of observations within studies is not large enough. In this case, the likelihood contribution of \( \theta_i^\beta \) can strongly deviate from the normal shape assumed in 1).

However, as we explain in Section 2.5, in this work we use Markov chain Monte Carlo (MCMC) simulations to approximate the posterior distributions of the model’s parameters. Therefore, the model presented does not require an approximated normal likelihood of \( \theta_i^\beta \). For example, if the reported results are given in two by two tables, where \( y_{i,1} \) is the number of events arising from \( n_{i,1} \) patients in the control group and \( y_{i,2} \) and \( n_{i,2} \) are the number of events and the number of patients in the treatment group, then the exact binomial likelihood is

\[
y_{i,1} \sim \text{Binomial}(p_{i,1},n_{i,1}), \quad \text{and} \quad y_{i,2} \sim \text{Binomial}(p_{i,2},n_{i,2}).
\]
The biased random effect of $\theta^B_i$ is defined as follows:

$$\theta^B_i = g(p_{1,2}) - g(p_{1,1}),$$

(13)

where $g(\cdot)$ is the link function (e.g., logit, probit and cloglog), and we follow the bias representation (2). In the similar way, if the outcome variable $y_i$ is modeled as Poisson or negative-binomial, the exact likelihood contribution can be applied.

### 2.5 Bayesian computations

The posterior distributions of the model parameters presented in Section 2 are calculated with MCMC simulations. We implemented these computations in JAGS (Plummer, 2003) and R.

In the Supporting Information, we provide the R script to produce the figures and the numerical results of Section 3. The statistical analysis of Section 3 is based on four parallel MCMC chains, with a length of 50,000 iterations. In each chain, we took the first 20,000 iterations as burn-in period and from the remaining 30,000 iterations we took one value every four iterations.

### 3 REAL DATA EXAMPLES

#### 3.1 Generalized evidence synthesis of pneumococcal polysaccharide vaccine against IPD

This case study, which has been presented in Section 1, is a recent meta-analysis performed by Falkenhorst et al. (2017). In this meta-analysis, the authors investigated the available evidence related to the effectiveness of pneumococcal polysaccharide vaccine (PPV23) against invasive IPD.

One characteristic of this meta-analysis is the inclusion of disparate data. From 17 studies, four were RCTs, five cohort studies, three case-control, and five case-case studies. One RCT (Alfageme et al. 2006) was excluded (for reporting no events in each treatment group), and the case-case studies did not report results of IPD. For the remaining 11 studies (three RCTs, five cohorts, and three case controls), the authors performed three meta-analyses by study types.

From the statistical point of view, a separate analysis for each study design has the assumption that the variability between study types is so large that we cannot expect to gain any information by combining study results. At the other extreme is the assumption of ignoring study designs and combining results in a single meta-analysis. We can expect that neither ignoring study types nor assuming that results are extremely different is the best way to synthesize this disparate evidence.

We applied the BC model to this meta-analysis using the following model’s specification: an approximate normal likelihood of $\theta^B_i$; we assumed that the bias direction was to the right; priors for location parameters were $\mu \sim \text{Normal}(0, 100)$, $\mu^B \sim \text{Uniform}(0, 10)$, prior for $\tau \sim \text{HC}(0.5)$.

In this meta-analysis, we have the number of studies of each study type as prior information. We use this information into the prior for $\pi \sim \text{Beta}(a_0, a_1)$, where $a_0 = 26.83$ and $a_1 = 15.48$. We used as the value of $\nu = 0.5$.

Figure 3 shows the resulting posterior distributions for $\mu$ and $\mu_{\text{biased}}$ after applying the BC model. The effective number of studies in this group is $N^+ = 4.13$, four studies compared to three RCTs. Therefore, by combining different study types we have added the contribution of one study into the BC group.

As a comparison, the Bayesian meta-analysis including only RCTs has a posterior mean and 95% interval of $-1.34 [-2.78, 0.10]$, which indicates that based on the RCTs’ evidence, the PPV23 is weakly effective against IPD. The posterior mean and 95% interval based on the BC model resulted in $-1.35 [-3.42, -0.76]$, which indicates effectiveness by combining RCT and BC OS evidence. By comparing the length of these two intervals, we can see that there is around 10.82% of shrinkage in the BC interval. Therefore, in this application, by including evidence from different study types we have increased the precision of inference.
However, the price of this bias correction has been the application of an informative prior for $\pi_{bias}$. We can display this effect in Figure 4 where the joint posterior distribution of $\mu_{bias}$ and $\pi_{bias}$ are presented. The contours are parallel to the $\mu_{bias}$ axes, which indicate that it is not possible to associate the probability of bias with the direction. The dashed line, which is the threshold of a single mode, shows that two modes are very likely in this meta-analysis.

The conclusion about bias depends strongly on the prior distribution because of the small sample size. However, if a prior bias between RCTs and OS is very likely, then it is worth paying the price and use an informative prior for $\pi_{bias}$.

### 3.2 Heart disease and bone marrow stem cells treatment

This case study corresponds to a meta-analysis of 31 RCTs of heart disease patients. In each trial, patients in the treatment group received bone marrow stem cells and patients in the control group received a placebo treatment (Nowbar et al., 2014). The treatment effect is defined as the difference of the ejection fraction between groups' means, which measures the improvement of the left ventricular function of the heart. The data of this meta-analysis are presented in Table 2.

Treatments based on bone marrow stem cells for patients with heart disease have been an exciting area of clinical research. However, trials’ results have been conflicting and it has been discovered that some early trials had unexpected efficacy results casting doubt on their validity (Francis, Mielewczik, Zargaran, & Cole, 2013). Further investigation has found evidence of scientific misconduct in papers reporting some trials’ findings (Abbott, 2014). Therefore, analyzing the all available evidence presented in Table 2 without incorporating a correction of internal validity bias can clearly lead to misleading results.

Suppose that for a while, we ignore the reporting issues mentioned above and we fit a Bayesian meta-analysis without internal validity bias-correction. The pooled treatment effect $\mu$ has a posterior mean and 95% posterior interval of 3.15 [1.54, 4.75], which is numerically similar to a classical random-effect meta-analysis. Can we conclude that we have enough evidence to demonstrate the efficacy of the treatment?

Unfortunately, these apparently confirming results are completely misleading. A warning comes from the very large heterogeneity between studies, summarized in the between-studies dispersion $\tau$ with a posterior mean and 95% interval of 3.88 [2.81, 5.30]. This results in a wide posterior predictive interval of $\mu$ of $[-4.86, 11.10]$ covering the no treatment effect.
We applied the BC meta-analysis model to the data of Table 2 using the following model’s specification: an approximate normal likelihood of $\theta_B$; we assume a bias direction to the right; priors for location parameters are $\mu \sim \text{Normal}(0, 100)$, $\mu_\beta \sim \text{Uniform}(0, 30)$, and the prior for $\tau \sim \text{HC}(0.5)$.

In this meta-analysis, it is not evident how to incorporate prior information for the probability of bias. Therefore, we use a noninformative prior for $\pi_{\text{bias}} \sim \text{Beta}(1, 1)$. The idea is to assess if we can learn the effect of bias from the data alone.

Figure 5 summarizes the results of the BC model. The BC component points out that there is a weak treatment effect. The posterior mean of the BC distribution is 1.22, and the 95% posterior interval is $[-0.75, 3.30]$.

The joint posterior distribution of the mean bias $\mu_\beta$ and the probability of bias $\pi_{\text{bias}}$ is displayed in Figure 6. In this example, we used noninformative priors for these two parameters. We can see in the contours that we have learned from the data. In addition, the posterior probability of having a bias component is 0.91, indicating the presence of studies with low internal validity bias.

In order to explain the sources of heterogeneity, Nowbar et al. (2014) investigated whether detected discrepancies in published trials could account for the variation in reported effect sizes. A discrepancy refers to two or more reported facts that cannot both be true because they are logically or arithmetically incompatible. For example, a discrepancy in baseline characteristics refers to sample or subgroup sizes that could not be an integer number of patients; a discrepancy in results refers to conflicts between tables and figures or impossible values.

The last column of Table 2 presents the total number of discrepancies that the reviewers found in each paper. We can see that there are three papers without discrepancies whereas the number of discrepancies in the other papers reached up to 55.

Figure 7 contrasts the bias-correction effect at the level of two studies, where the dashed lines correspond to the likelihoods of the $\theta_{\text{bias}}$ and the solid lines to the BC posteriors of $\theta$. On the left panel, we have a study without discrepancies. In this case, the BC posterior slightly corrected the dispersion. On the right panel, a notable correction of bias was applied.
FIGURE 5  Results of the BC approach to a meta-analysis of studies applying treatments based on bone marrow stem cells to improve the left ventricular function. Posteriors of the pooled treatment effect of BC and the biased component

FIGURE 6  Diagnostic plot for the BC model: Joint posterior distribution of the mean bias and probability of bias. Noninformative priors have been applied. The probability contours show the effect of learning from the study results. The scatter plot corresponds to random samples from the MCMC iterations. The horizontal dashed line is the threshold of a single mode, and most of the posterior density of bias are above this value, which indicates the existence of two mixture components.
TABLE 2  Results from 31 RCTs of heart disease patients, where the treatment group received bone marrow stem cells and the control group a placebo treatment. The treatment effect is measured as the difference of the ejection fraction between groups. Discrepancies refer to two or more reported facts that cannot be true because they are logically or arithmetically incompatible. The source of this data is the meta-analysis performed by Nowbar et al. (2014)

| Author                  | Year | Treatment effect | SE  | Sample size | Number of discrepancies |
|-------------------------|------|------------------|-----|-------------|------------------------|
| Quyyumi                 | 2011 | 1.50             | 3.67| 21          | 17                     |
| Lunde                   | 2007 | 1.10             | 2.09| 100         | 7                      |
| Srimahachota            | 2011 | −1.70            | 2.91| 23          | 7                      |
| Meyer                   | 2006 | 0.80             | 2.78| 60          | 4                      |
| Meluzín                 | 2006 | 7.00             | 0.63| 40          | 4                      |
| Piepoli                 | 2010 | 7.80             | 2.76| 38          | 21                     |
| Suárez de Lezo          | 2007 | 14.00            | 4.05| 20          | 13                     |
| Huikuri HV              | 2008 | 5.40             | 2.44| 77          | 18                     |
| Perin                   | 2012 | 2.70             | 1.20| 82          | 16                     |
| Assmus                  | 2006 | 4.10             | 0.98| 46          | 0                      |
| Assmus                  | 2013 | 2.20             | 0.88| 79          | 27                     |
| Assmus                  | 2010 | 2.50             | 1.01| 187         | 11                     |
| Hendrikx                | 2006 | 2.50             | 3.96| 20          | 2                      |
| Hirch                   | 2011 | −0.20            | 1.17| 127         | 3                      |
| Perin                   | 2012 | −2.00            | 2.09| 20          | 0                      |
| Traverse                | 2011 | −3.00            | 2.03| 81          | 0                      |
| Ang                     | 2008 | 3.60             | 2.52| 17          | 2                      |
| Rodrigo                 | 2012 | 4.00             | 1.28| 40          | 3                      |
| Herbolts                | 2009 | −1.50            | 1.75| 66          | 8                      |
| Castellani              | 2010 | 8.80             | 6.07| 10          | 6                      |
| Maureira                | 2012 | 4.00             | 3.89| 14          | 2                      |
| Ribero dos Santos       | 2012 | −0.20            | 1.54| 183         | 19                     |
| Traverse                | 2010 | −3.20            | 3.63| 40          | 7                      |
| Patel                   | 2004 | 10.10            | 1.21| 20          | 11                     |
| Tse                     | 2007 | 5.40             | 2.54| 27          | 15                     |
| Cao                     | 2009 | 3.50             | 1.04| 86          | 6                      |
| Sürder                  | 2013 | 1.25             | 1.56| 118         | 9                      |
| Ge                      | 2006 | 6.70             | 3.16| 20          | 2                      |
| Traverse                | 2012 | −0.10            | 2.03| 112         | 2                      |
| Wöhrlie                 | 2010 | −3.90            | 2.62| 40          | 1                      |
| Yousef (Strauer)        | 2009 | 10.40            | 1.01| 116         | 55                     |

We applied the model extension presented in Section 2.3 using a logistic link function. This analysis allows us to explore the relationship between the probability of bias and the number of studies’ discrepancies. Figure 8 displays the posterior mean and 95% intervals of the probability of bias as a function of discrepancies. The points correspond to the posterior probability of bias for each study, where the diameter of each point is proportional to the number of discrepancies in the study. The posterior mean and 95% interval of $\alpha_1$ are $0.132 \ [-0.019, 0.314]$, with a posterior probability that $\alpha_1 > 0$ of 0.955. These results indicate that the probability of bias increases with the number of discrepancies. Therefore, studies that had published exaggerated treatment effects containing a large number of discrepancies are at high risk of bias.
**FIGURE 7** Comparison between the likelihood of $\theta^p$ (dashed lines) and the BC $\theta$ (solid line). The left panel corresponds to a study without reported discrepancies, and the right panel corresponds to a study with 55 discrepancies.

**FIGURE 8** Results of the meta-regression approach based on a logistic link function: Posterior distribution of the relationship between the number of discrepancies and the probability of bias. The points correspond to the posterior probability of bias for each study, where the diameters are proportional to the number of discrepancies.
Meta-analysis methods allow researchers to combine results from multiple pieces of evidence into a single analysis. These techniques are used to extend the scope of a single study by combining results from several studies. In particular, a meta-analysis of RCTs, addressing the same primary research question, is considered the gold standard in clinical evidence.

However, there are important reasons to include OS in a meta-analysis. In particular, RCTs’ results might not be representative of real-world populations. Consequently, we need meta-analysis methods to combine disparate pieces of evidence to gain information in order to answer such questions. Moreover, the amount of evidence in a particular problem could be very limited and by including OS we can complement results of RCTs. As we have seen in this paper, in order to incorporate observational evidence or evidence with varying quality in meta-analysis, we need to adjust for internal validity bias.

The problem is that we can only learn about internal validity bias indirectly, by what has been reported in a study’s publication. Therefore, the adjustment of this bias is a subtle task. In this paper, we have approached this problem by using a random effect model that explicitly accounts for this bias. This model has been elaborated for the syntheses of different study types, where the presence of internal validity bias between different study designs is evident. We showed that the model could also be applied as a generic Bayesian meta-analysis model for RCTs.

We have applied some rather strong assumptions in order to make the model parameters identifiable from the data. For example, we assumed that \( \mu_{\beta_i} = \mu_\beta \). One can argue that if study level covariates that explain study conduct are available, then these covariates may be incorporated in a meta-regression equation and avoid the assumption that \( \mu_{\beta_i} = \mu_\beta \). However, this kind of analysis may not always be possible or appropriate. In particular, this is because the design and conduct are assessed by the report of the study and not from the study itself. For this reason, the modeling strategy was to handle study bias indirectly at the price of making strong assumptions.

As highlighted by one of the referees, it may be of concern that we are only modeling “outliers” or “large treatment effects” and not internal validity bias. I prefer to cast this issue by understanding if the data at hand represent a direct outcome, for example, individuals’ data in a single RCT, or whether what we have is an indirect evidence. Outliers’ detection refers to direct modeling of data, where an extreme observation can destabilize the conclusions of a statistical analysis. In meta-analysis, however, we are synthesizing indirect evidence, where results represent one possibility of many. A study that published a “large treatment effect” with poor quality or low internal validity due to study’s design represents “a conflict” in the body of evidence. This kind of interpretation is the hallmark of meta-analysis which models indirect results.

What we are trying to achieve in the model presented in Section 2 is to cluster studies’ results that are in conflict with the other results and to down-weight their influence in the meta-analysis. Although the aim is to do this automatically, without the interpretation of results and without making a sensitivity analysis regarding the direction and the probability of bias, we cannot decide whether to combine or not to combine observational and experimental evidence in a meta-analysis.

There are several possible extensions and applications that we did not cover in this paper. These extensions include the network meta-analysis of randomized and OS, the meta-analysis of individual participant data, and the meta-analysis of diagnostic and prognostic studies. These meta-analyses often include studies of varying quality and designs.

We have two important conclusions in this work. The first one is about the use of flexible random-effects distributions in a meta-analysis (e.g., finite mixtures, Bayesian nonparametric). We have learned that it makes sense to extend the flexibility of the random effects if we can identify the cluster of high-quality evidence. If this is not possible, these flexible models may not add value to the meta-analysis. The second conclusion of this paper is that ignoring internal validity bias in meta-analysis is a form of model misspecification that may result in misleading conclusions. The approach presented in this paper is a promising solution to this important problem in meta-analysis.

ACKNOWLEDGMENTS
This work was presented by Pablo Emilio Verde at the 40th Annual Conference of the International Society for Clinical Biostatistics in Leuven 2019 (ISCB 2019). The author thanks Professor Emmanuel Lesaffre and Professor Stephen Senn for a fruitful exchange of ideas at ISCB 2019. I am thankful to Dr. Daniel Curcio for introducing me to the meta-analysis of IPD and to Professor Christian Ohmann who pointed out the stem cells meta-analysis. I also thank two reviewers, who helped to greatly improve the quality of this paper. The author was partially supported by the German Research Foundation project DFG VE 896/1-1.

Open access funding enabled and organized by Projekt DEAL.
CONFLICT OF INTEREST
The author has declared no conflict of interest.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are openly available in my R package jarbes in CRAN at https://cran.r-project.org/web/packages/jarbes/index.html.

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

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

ORCID
Pablo Emilio Verde https://orcid.org/0000-0001-8625-7242

REFERENCES
Abbott, A. (2014). Evidence of misconduct found against cardiologist, Nature. News Blog. http://blogs.nature.com/news/2014/02/evidence-of-misconduct-cardiologist.html
Dias, S., Welton, N. J., Marinho, V. C. C., Salanti, G., Higgins, J. P. T., & Ades, A. E. (2010). Estimation and adjustment of bias in randomized evidence by using mixed treatment comparison meta-analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society), 173, 613–629.
Dominguez, A., Salleras, L., Fedson, D. S., Izquierdo, C., Ruiz, L., Ciruela, P., … Casal, J. (2005). Effectiveness of pneumococcal vaccination for elderly people in catalonia, Spain: A case-control study. Clinical Infectious Diseases, 40, 1250–1257.
Droitcour, J., Silberman, G., & Chelimsky, E. (1993). Cross-design synthesis: A new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. International Journal of Technology Assessment in Health Care, 9, 440–449.
Eddy, D. M., Hasselblad, V., & Schachter, R. (1992). Meta-analysis by the confidence profile method: The statistical synthesis of evidence. San Diego, CA: Academic Press.
Efthimiou, O., Mavridis, D., Debray, T., Samara, M., Belger, M., Siontis, G., … Salanti, G. (2017). Combining randomized and non-randomized evidence in network meta-analysis. Statistics in Medicine, 36, 1210–1226.
Falkenhorst, G., Remschmidt, C., Harder, T., Hummers-Pradier, E., Wichmann, O., & Bogdan, C. (2017). Effectiveness of the 23-valent pneumococcal polysaccharide vaccine (PPV23) against pneumococcal disease in the elderly: Systematic review and meta-analysis. PLoS ONE, 12(1), e0169368. https://doi.org/10.1371/journal.pone.0169368.
Ferguson, J., Alvarez-Iglesias, A., Newell, J., Hinde, J., & O’Donnell, M. (2019). Joint incorporation of randomized and observational evidence in estimating treatment effects. Statistical Methods in Medical Research, 28(1), 235–247.
Francis, D. P., Mielewicz, M., Zargaran, D., & Cole, G. D. (2013). Autologous bone marrow-derived stem cell therapy in heart disease: Discrances and contradictions. International Journal of Cardiology, 168, 3381–403.
Greenhouse, J., Anderson, H., Bridge, J., Libby, A., Valuck, R., & Kelleher, K. (2017). Combining information from multiple data sources: An introduction to cross-design synthesis with a case study. In C. Gatsionis & S.C. Morton (Eds.), Methods in comparative effectiveness research (pp. 203–225). Boca Raton, FL: Chapman & Hall/CRC Press.
Hechter, R. C., Chao, C., Jacobsen, S. J., Slezak, J. M., Quinn, V. P., Van Den Eeden, S. K., Tseng, H. F. (2012). Clinical effectiveness of pneumococcal polysaccharide vaccine in men: California Men’s Health Study. Vaccine, 30, 5625–5630.
Henderson, N. C., Varadhan, R., & Weiss, C. O. (2017). Cross-design synthesis for extending the applicability of trial evidence when treatment effect is heterogenous: Part II. Application and external validation. Communications in Statistics: Case Studies, Data Analysis and Applications, 3, 7–20.
Honkanen, P. O., Keistinen, T., Miettinen, L., Herva, E., Sankilampi, U., Lääri, E., … Mäkelä, P. H. (1999). Incremental effectiveness of pneumococcal vaccine on simultaneously administered influenza vaccine in preventing pneumonia and pneumococcal pneumonia among persons aged 65 years or older. Vaccine, 17, 2493–2500.
Ibrahim, J., Chen, M., Gwon, Y., & Chen, F. (2015). The power prior: Theory and applications. Stat Med, 34(28), 3724–3749.
Jackson, D., & White, I. R. (2018). When should meta-analysis avoid making hidden normality assumptions? Biometrical Journal, 60, 1040–1058.
Jackson, L. A., Neuzil, K. M., Yu, O., Benson, P., Barlow, W. E., Adams, A. L., … Thompson, W. W.; for the Vaccine Safety Datalink. (2003). Effectiveness of pneumococcal polysaccharide vaccine in older adults. New England Journal of Medicine, 348, 1747–1755.
Kaizar, E. (2011). Estimating treatment effect via simple cross design synthesis. Statistics in Medicine, 130(25), 2986–3009.
Lange, K. and Sinusheimer, J. S. (1993). Normal/Independent Distributions and Their Applications in Robust Regression. Journal of Computational and Graphical Statistics, 2, 175–198.
Leventer-Roberts, M., Feldman, B. S., Brufman, I., Cohen-Stavi, C. J., Hoshen, M., Balicer, R. D. (2015). Effectiveness of 23-valent pneumococcal polysaccharide vaccine against invasive disease and hospital-treated pneumonia among people aged ≥65 years: A retrospective case-control study. Clinical Infectious Diseases, 60, 1472–1480.
How to cite this article: Verde PE. A bias-corrected meta-analysis model for combining, studies of different types and quality. *Biometrical Journal*. 2021;63:406–422. https://doi.org/10.1002/bimj.201900376