

# A Generalized Inferential Model for Meta-Analyses Based on Few Studies

Joyce Cahoon and Ryan Martin<sup>1</sup>

<sup>1</sup>*Department of Statistics  
North Carolina State University, Raleigh, North Carolina, USA*

Received: 28 February 2020; Revised: 13 August 2020; Accepted: 01 September 2020

---

## Abstract

Meta-analysis based on only a few studies remains a challenging problem, as an accurate estimate of the between-study variance is apparently needed, but hard to attain, within this setting. Here we offer a new approach, based on the *generalized inferential model* framework, whose success is based on marginalizing out the between-study variance, so that an accurate estimate is not essential. We show theoretically that the proposed solution is at least approximately valid, with numerical results suggesting it is, in fact, nearly exact. We also demonstrate that the proposed solution outperforms existing methods across a wide range of scenarios.

*Key words:* Confidence interval; Monte Carlo; Normal random effects model; Plausibility function; Profile likelihood.

**AMS Subject Classifications:** 62F25, 62J05, 62F86

---

## 1. Introduction

The most important scientific questions are likely to be pursued by multiple researchers, resulting in separate analyses that may be combined via a single *meta-analysis* to attain stronger and more definitive conclusions. But even when it is appropriate to combine the results from multiple studies, there is often a non-negligible amount of between-study heterogeneity, which is difficult to estimate accurately when the number of studies for meta-analysis is small. Interestingly, meta-analyses with as few as three studies are the norm, not the exception (Davey *et al.* 2011), so there is considerable interest in developing improved methods for inference in this setting of combining results from just a few heterogeneous studies.

To set the scene, consider the classical normal–normal random-effects model where each study  $k$  included for meta-analysis provides data  $(Y_k, \sigma_k^2)$ . These are modeled as

$$(Y_k | M_k) \stackrel{\text{iid}}{\sim} \mathbf{N}(M_k, \sigma_k^2), \quad M_k \stackrel{\text{iid}}{\sim} \mathbf{N}(\mu, \nu), \quad k = 1, \dots, K. \quad (1)$$

Here,  $M_k$  denotes the random effect from study  $k$ ,  $\sigma_k^2 > 0$  the variance within study  $k$ ,  $\mu$  the underlying population effect and  $\nu > 0$  the between-study variance. Note that taking the  $\sigma_k^2$ 's as part of the observed data implies that these values are *known*, a common assumption

in this literature (*e.g.*, DerSimonian and Laird 1986; Michael *et al.* 2019). This hierarchical formulation of the model indicates that the  $K$  studies have something in common, namely, a tendency toward  $\mu$ , but that they are not fully homogeneous and that this degree of heterogeneity, controlled by  $\nu$ , is unknown. Therefore, the goal is inference on the unknown mean  $\mu$ , with  $\nu$  as an unknown nuisance parameter.

In the simple case where  $\nu$  is *known*, the meta-analysis is straightforward. Marginally,  $Y_k$  are independent, distributed as  $\mathbf{N}(\mu, \sigma_k^2 + \nu)$ , for  $k = 1, \dots, K$ . It is easy to check that the minimum variance unbiased estimator of  $\mu$  is

$$\hat{\mu}(\nu) = \frac{\sum_{k=1}^K w_k(\nu) Y_k}{\sum_{k=1}^K w_k(\nu)}, \quad \text{where } w_k(\nu) = (\sigma_k^2 + \nu)^{-1},$$

and its variance is  $\{\sum_{k=1}^K w_k(\nu)\}^{-1}$ . Of course, the more common scenario in applications is that  $\nu$  is *unknown* and, somehow, the data must be used to account for that additional uncertainty. A first idea is to estimate  $\nu$  by some function  $\hat{\nu}$  of the data, and then plug that into the formula for the estimator of  $\mu$ , *i.e.*,  $\hat{\mu}(\hat{\nu})$ . A well-known strategy is that proposed by DerSimonian and Laird (1986), but there are others, *e.g.*, Paule and Mandel (1982) and Cochran (1954). Unfortunately, the number of studies,  $K$ , to be combined is often relatively small (Davey *et al.* 2011), say,  $K \leq 7$ , and obtaining reliable estimates of the between-study variance based on so few samples is a challenge. Besides, this plug-in style procedure does not naturally account for uncertainty in  $\nu$ , so any inference drawn can only be (provably) valid in an asymptotic ( $K \rightarrow \infty$ ) sense, which may not be especially meaningful in applications where  $K \leq 7$ . More tangibly, confidence intervals based on the DerSimonian–Laird plug-in style method have shown to perform poorly even when  $K$  is as large as 20 (Liu *et al.* 2018); see, also, Viechtbauer (2005), DerSimonian and Kacker (2007), Sidik and Jonkman (2007), Jackson *et al.* (2010), Chung *et al.* (2013), and Veroniki *et al.* (2016). Therefore, there is a need for methods that marginalize out the nuisance parameter,  $\nu$ , and achieve frequentist error rate control even when the number of studies is small.

To alleviate this problem of estimating  $\nu$ , likelihood methods have been proposed, *e.g.*, the log likelihood ratio in Goodman (1989), the profile likelihood in Hardy and Thompson (1996), and the signed profile log likelihood in Severini (2000), among others. Unlike previous plug-in methods, these likelihood based procedures introduced an appropriate widening in confidence intervals to deal with the imprecision in estimating  $\nu$ . In fact, Guolo (2012) was able to improve upon these first-order inference results by introducing a Skovgaard correction to the signed profile log likelihood, making it asymptotically standard normal with an error that vanishes at a rate of  $n^{-3/2}$ ,  $n \rightarrow \infty$ . Resorting to higher-order asymptotics with the Skovgaard correction, however, still failed to modify the likelihood to the extent necessary to achieve the nominal coverage probability when the number of studies available is small.

So-called *exact methods*, featuring such frequentist guarantees, have been offered as alternatives to the plug-in style and likelihood methods described above. These include the methods in Follmann and Proschan (1999), Liu *et al.* (2018), and Wang and Tian (2018) which, in one way or another, are based on permutation distributions. While these permutation methods can produce confidence intervals that achieve nominal frequentist coverage, the discreteness of the permutation distribution makes the results overly conservative, unless  $K$  is relatively large. To our knowledge, the most recent work on exact methods for

meta-analysis is Michael *et al.* (2019), and since this shares a number of similarities to our proposal, at least in terms of its construction, we describe this in detail in Section 3.1.

Given that the goal is to develop a method for meta-analysis that controls Type I error, even when  $K$  is small, it makes sense to consider the one general, normative framework we are aware of that offers such guarantees. Specifically, Martin and Liu (2013) present a construction of what they call an *inferential model* that leads to provably valid inference, no asymptotic justification required; see Martin and Liu (2016) for a monograph-length introduction, and Martin (2019) for a survey of some recent developments. The distinguishing feature of this approach is the user-specified random set that leads to a sort of “posterior” (but *not* a probability measure) on the parameter space. The distribution of this random set can then be used to visualize the information data provides about the parameter of interest as well as construct inference procedures. We will briefly review the basic construction and its properties in Section 2.1. It may happen that the basic inferential model construction is difficult to carry out in an application, so Martin (2015, 2018) developed a simpler and more direct *generalized* version. Since this is the strategy we follow here for meta-analysis, we provide a brief review in Section 2.2.

This meta-analysis application boils down to a marginal inference problem, *i.e.*,  $\nu$  is an unknown nuisance parameter to be marginalized out so that we can make inference about  $\mu$ . The generalized inferential model framework provides at least two strategies for marginalization, and we will show that the method proposed in Michael *et al.* (2019) is itself a generalized inferential model based on one particular choice of marginalization strategy. Our proposed method, on the other hand, is based on a different and arguably more natural choice of marginalization strategy, leading to a method that performs better than theirs in a variety of respects. In Section 3, we describe the construction of a generalized inferential model for meta-analysis, show that the solution in Michael *et al.* (2019) is a special case, and present our proposed method. Details about the computation and theoretical justification are also provided. Numerical comparisons in simulated data experiments demonstrate that our proposed method outperforms existing methods in terms of both validity and efficiency across a broad range of scenarios. Two real applications are presented in Section 5 and some concluding remarks are given in Section 6.

## 2. Background

### 2.1. Basic inferential models

Fisher and later Dempster aimed to develop a framework of probabilistic inference without prior distributions, *i.e.*, a prior-free alternative to Bayesian inference. These approaches, however, failed to reach the statistical mainstream, largely because the derived procedures have no frequentist guarantees. To fill that gap, Martin and Liu (2013) argued that frequentist guarantees could be achieved by supplementing the structural, pivotal, or functional model formulation of Fisher (1956), Dempster (2008), Fraser (1968), Barnard (1995), Dawid and Stone (1982), Taraldsen and Lindqvist (2013), and others, with an appropriate user-specified random set. As a consequence of this use of a random set, the inferential output is described by a (data-dependent) non-additive plausibility function (*e.g.*, Shafer 1976) instead of an additive Bayesian, fiducial, or structural posterior probability distribution. The added complexity of non-additivity is not for its own sake, however, it is actually necessary (Balch

*et al.* 2019) for the strong validity property in (3) that gives the inferential model's belief assignments a meaningful scale for interpretation and, furthermore, leads to the frequentist error rate control guarantees.

In general, we start with a sampling model,  $Y \sim P_{Y|\theta}$ , for the observable data  $Y$ , depending on some unknown parameter  $\theta \in \Theta$ . As one would if the goal were to simulate, we express the model as

$$Y = a(\theta, U), \quad U \sim P_U. \quad (2)$$

where  $a$  is a known function, and  $P_U$  is a known distribution on the space  $\mathbb{U}$ . We call the relationship in (2) an *association* between observable data  $Y$ , unknown parameter  $\theta$ , and auxiliary variable  $U$ . This association step is the starting point in the construction of a valid inferential model.

*A-step.* Specify an association of the form (2) and then define the set

$$\Theta_y(u) = \{\theta : y = a(\theta, u)\}, \quad u \in \mathbb{U}.$$

*P-step.* Specify a predictive random set  $\mathcal{S}$ , taking values in the power set of  $\mathbb{U}$ , whose contour function,  $f(u) = P_{\mathcal{S}}(\mathcal{S} \ni u)$ , is such that  $f(U) \sim \text{Unif}(0, 1)$ , when  $U \sim P_U$ . Remarks on the choice of random set  $\mathcal{S}$ —and the corresponding contour function  $f$ —are provided in Martin and Liu (2016, Ch. 4); see, in particular, Corollary 4.1.

*C-step.* Combine the ingredients in the A- and P-steps to get a new random set

$$\Theta_y(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \Theta_y(u),$$

and, for inference about  $\theta$ , return the distribution of this random set summarized by its plausibility function

$$\text{pl}_y(A) = P_{\mathcal{S}}\{\Theta_y(\mathcal{S}) \cap A \neq \emptyset\}, \quad A \subseteq \Theta.$$

The distribution of this random set is interpreted as a measure of how *plausible* the hypothesis “ $\theta \in A$ ” is, based on data  $y$  and the posited model.

The most unique feature of this construction is the random set  $\mathcal{S}$ . In our meta-analysis problem, specification of the random set is straightforward as outlined in Section 2.2, but the general details can also be found in Martin and Liu (2016). What matters is that the properties required of the random set make the inferential model *valid*, *i.e.*,

$$\sup_{\theta \in A} P_{Y|\theta}\{\text{pl}_Y(A) \leq \alpha\} \leq \alpha \quad \text{for all } \alpha \in (0, 1) \text{ and all } A \subseteq \Theta. \quad (3)$$

An important consequence of this validity property is the control it provides on the performance of statistical procedures derived from the inferential model output. Indeed, a test that rejects a hypothesis “ $\theta \in A$ ” if  $\text{pl}_y(A) \leq \alpha$  will obviously control the frequentist Type I error rate at level  $\alpha$ . Similarly, a  $100(1 - \alpha)\%$  plausibility region for  $\theta$ , given by

$$\{\theta : \text{pl}_y(\theta) > \alpha\}, \quad \text{where } \text{pl}_y(\theta) := \text{pl}_y(\{\theta\}),$$

has frequentist coverage probability of (at least)  $1 - \alpha$ . These properties are exact in the sense that they do not require any asymptotic approximations. The pointwise plausibility function,  $\theta \mapsto \text{pl}_y(\theta)$ , is also a useful visualization tool, not unlike a Bayesian posterior density function; see Figure 1 below.

## 2.2. Generalized inferential models

Since the inferential model construction and validity result is general, efficiency often becomes a concern as the dimension of the auxiliary domain increases with the size of data. To avoid the possible need to specify such a complex, high-dimensional random set, some non-trivial manipulations are required (*e.g.*, Martin and Liu 2015ab) that can be difficult to carry out in a given problem. This motivated Martin (2015, 2018) to develop a construction based on a more general formulation and establish conditions under which the corresponding inferential model is valid. An advantage of this generalized approach is that there is no need for the aforementioned manipulations, hence it is easier to apply.

A generalized inferential model begins with defining a real-valued function  $(y, \theta) \mapsto T_y(\theta)$ . When  $Y \sim \mathbf{P}_{Y|\theta}$ , the random variable  $T_Y(\theta)$  has a distribution, which we represent with  $G_\theta$ . The generalized association then extends the notion in Section 2.1 by connecting the data, parameter, and auxiliary variable via the expression

$$T_Y(\theta) = G_\theta^{-1}(U), \quad U \sim \text{Unif}(0, 1).$$

In this paper, we will assume (without loss of generality) that  $T_y(\theta)$  is large when data  $y$  and parameter value  $\theta$  disagree; therefore,  $T_y(\theta)$  is a sort of distance between data  $y$  and parameter  $\theta$ . The first step to our generalized inferential model (A-step) then yields

$$\Theta_y(u) = \{\theta : G_\theta(T_y(\theta)) = u\}, \quad u \in (0, 1).$$

The P-step, as before, requires the introduction of some random set in the  $u$ -space; but the structure that has been imposed here virtually determines it. We thus take

$$\mathcal{S} = [0, \tilde{U}], \quad \tilde{U} \sim \text{Unif}(0, 1).$$

The C-step returns a new random set

$$\Theta_y(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \Theta_y(u) = \{\theta : G_\theta(T_y(\theta)) \leq \tilde{U}\}, \quad \tilde{U} \sim \text{Unif}(0, 1).$$

Note that this set contains those parameter values that agree with  $y$  to some degree, and that this degree is calibrated so that validity holds. That is, if

$$\text{pl}_y(\theta) = \mathbf{P}_{\mathcal{S}}\{\Theta_y(\mathcal{S}) \ni \theta\} = \mathbf{P}_{\tilde{U}}\{G_\theta(T_Y(\theta)) \leq \tilde{U}\} = 1 - G_\theta(T_y(\theta)),$$

then we immediately see that  $\text{pl}_Y(\theta) \sim \text{Unif}(0, 1)$  under  $Y \sim \mathbf{P}_{Y|\theta}$  and, hence, the validity property as stated in (3) holds.

The approach described above returns a plausibility function defined on the full parameter space,  $\Theta$ . From this, one can carry out marginal inference on any feature,  $\psi = \psi(\theta)$ , of  $\theta$  via optimization. In particular, following Shafer (1987, Sec. G), the corresponding marginal point plausibility function for  $\psi$  is

$$\text{mpl}_y(\psi) = \sup_{\theta: \psi(\theta) = \psi} \text{pl}_y(\theta), \tag{4}$$

and the validity properties associated with  $\text{pl}$  carry over immediately to  $\text{mpl}$ .

But there are cases, like the one we consider in Section 3, where interest is exclusively in a specific feature of  $\theta$ , and it is beneficial to construct a marginal inferential model directly. Express the full parameter as  $\theta = (\psi, \eta)$ , where  $\psi$  and  $\eta$  are the interest and nuisance parameters, respectively. Next, define a function  $(y, \psi) \mapsto T_y(\psi)$  that only directly involves the interest parameter, again, with the property that large values of the function correspond to cases where data and the interest parameter disagree. An example of such a function is the negative log relative profile likelihood, as in (6). If the distribution of  $T_Y(\psi)$ , as a function of  $Y \sim P_{Y|\psi, \eta}$ , does not depend on  $\eta$ , then construction of the generalized inferential model for  $\psi$  proceeds exactly as above, with any fixed value of  $\eta$ . This is very similar to the developments of Tsui and Weerahandi (1989) in the framework of generalized p-values. In many applications, however, including our meta-analysis problem, the distribution of  $T_Y(\psi)$  does depend on the nuisance parameter, so some non-trivial adjustments are required. We discuss this in detail in Section 3.

### 3. Inferential Models for Meta-Analysis

#### 3.1. Construction

For our meta-analysis case, write  $P_{Y|\mu, \nu}$  for the joint distribution of  $Y = (Y_1, \dots, Y_K)$ , where  $Y_k$  are independently generated from a  $N(\mu, \sigma_k^2 + \nu)$ . It is straightforward to write down an association that links the data  $Y$ , the unknown parameter  $\theta = (\mu, \nu)$ , and a set of auxiliary variables, *e.g.*,

$$Y_k = \mu + (\sigma_k^2 + \nu)^{1/2}U_k, \quad k = 1, \dots, K,$$

where  $U_k$ 's are iid  $N(0, 1)$ . Following Martin and Liu (2015a), the next step would be to reduce the dimension of  $(U_1, \dots, U_K)$  to match that of  $\theta$ . This step turns out to be challenging but, fortunately, a generalized inferential model is within reach.

The full parameter is  $\theta = (\mu, \nu)$  but, since only  $\mu$  is of interest, marginalization is desired. As discussed in Section 2.2, there are at least two ways to proceed. The first is to start with a summary  $T_Y(\theta) = T_Y(\mu, \nu)$  of the data and full parameter, which takes large values when data and the candidate parameters disagree, and then marginalize to the  $\mu$ -space *after* constructing the plausibility function on the  $(\mu, \nu)$ -space. That is, we define a plausibility function on the full parameter space as

$$\text{pl}_y(\mu, \nu) = 1 - G_{\mu, \nu}(T_y(\mu, \nu)),$$

where  $G_{\mu, \nu}$  is the distribution of  $T_Y(\mu, \nu)$  under  $Y \sim P_{Y|\mu, \nu}$ . Like in (4), we obtain our desired marginal plausibility by optimization:

$$\text{mpl}_y(\mu) = \sup_{\nu} \text{pl}_y(\mu, \nu).$$

The corresponding  $100(1 - \alpha)\%$  plausibility interval for  $\mu$  is  $\{\mu : \text{mpl}_y(\mu) > \alpha\}$ , corresponding to a projection of the joint plausibility region for the full parameter onto the  $\mu$ -space. After some reflection on the solution in Michael *et al.* (2019), one sees that it is precisely a generalized inferential model as just described, with  $T_y(\mu, \nu)$  given by

$$T_y(\mu, \nu) = T_w(\mu) + c_0 \log \frac{L_y(\mu, \hat{\nu}_{DL})}{L_y(\mu, \nu)}, \quad (5)$$

a linear combination of DerSimonian and Laird’s Wald-type summary statistic  $T_w(\mu)$  and a log likelihood ratio, with a constant  $c_0$  controlling its contribution. Those authors do not describe their proposal as a (generalized) inferential model, but we believe that this perspective is beneficial both for developing some intuition about their solution and for comparing with our proposed solution.

Despite the ease of marginalizing out the nuisance parameter from the joint plausibility function, we adopt the second strategy for marginalization discussed in Section 2.2 that eliminates the nuisance parameter *before* constructing the plausibility function. That is, we start with a function,  $T_y(\mu)$ , that does not directly involve  $\nu$ . Like in Goodman (1989), we recommend the use of a negative log relative profile in which

$$T_y(\mu) = -\log \frac{\sup_{\nu} L_y(\mu, \nu)}{\sup_{\mu, \nu} L_y(\mu, \nu)}, \quad \mu \in \mathbb{R}. \quad (6)$$

Here  $L_y(\mu, \nu) \propto \prod_{k=1}^K (\sigma_k^2 + \nu)^{-1/2} \exp\{-\frac{1}{2}(\sigma_k^2 + \nu)^{-1}(y_k - \mu)^2\}$  is the likelihood function under the assumed model  $P_{Y|\mu, \nu}$ . Note that the  $\nu$  value at which the maximum is attained in the numerator—call it  $\hat{\nu}_{\mu}$ —depends on the specified value of  $\mu$ . As before, we define the distribution function of  $T_Y(\mu)$  under the model  $Y \sim P_{Y|\mu, \nu}$  as

$$G_{\nu}(t) = P_{Y|0, \nu}\{T_Y(0) \leq t\}, \quad t > 0.$$

Here we have inserted the default zero value for  $\mu$  because the location model structure means the distribution of  $T_Y(\mu)$  does not depend on the value of  $\mu$ , when  $Y \sim P_{Y|\mu, \nu}$ . Following the remainder of the construction outlined in Section 2.2, we arrive at a marginal point plausibility function

$$\text{“mpl}_y(\mu)\text{”} = 1 - G_{\nu}(T_y(\mu)).$$

The quotation marks on the left-hand side are to signal that this is not a function that we can actually work with because the right-hand side depends on the unknown value of the nuisance parameter  $\nu$ . To overcome this, we will use a plug-in estimate for  $\nu$ , where it appears in  $G_{\nu}$ . Before proceeding, it is important to emphasize that our plug-in proposal is fundamentally different than those mentioned in Section 1; we discuss this in more detail in Section 3.3. For our plug-in estimator, we propose to use that  $\nu$  value where the maximum in the numerator of the profile likelihood is attained, namely,  $\hat{\nu}_{\mu}$ , which implicitly depends on data  $y$ . Putting it all together, our proposed marginal point plausibility function for  $\mu$  is

$$\text{mpl}_y(\mu) = 1 - G_{\hat{\nu}_{\mu}}(T_y(\mu)), \quad \mu \in \mathbb{R}, \quad (7)$$

which is now just a function of data and the generic argument  $\mu$ . This function can be plotted to visualize what the data suggests about where the true value of  $\mu$  is and, more formally, we can read off a  $100(1 - \alpha)\%$  marginal plausibility interval for  $\mu$  as follows:

$$\{\mu : \text{mpl}_y(\mu) > \alpha\}.$$

Computation of  $\text{mpl}_y$  requires an approximation of the analytically intractable distribution  $G_{\hat{\nu}_{\mu}}$ , but this is straightforward to do via Monte Carlo; see Algorithm 1. And once  $\text{mpl}_y(\mu)$  is available on a grid of values, extracting the plausibility interval for  $\mu$  is easy, but the endpoints could be targeted more directly using, say, the proposed Monte Carlo method coupled with stochastic approximation.

**Algorithm 1:** Monte Carlo approximation of  $G_{\hat{\nu}_\mu}$ 


---

```

1 Generate  $M$  samples of  $K$  study-level errors  $e_{1m}^*, \dots, e_{Km}^* \sim \mathbf{N}(0, 1)$  Set a fine grid of
   $\mu$  values for each  $\mu$  value on the specified grid do
2   Find  $\hat{\nu}_\mu$  for the observed data  $y$  and given  $\mu$ ;
3   for  $m = 1, \dots, M$  do
4     Set  $Y_{km}^* = (\sigma_k^2 + \hat{\nu}_\mu)^{1/2} e_{km}^*$  for  $k = 1, \dots, K$ ;
5     Calculate  $T_m^* = T_{Y_m^*}(0)$  based on  $Y_m^* = (Y_{1m}^*, \dots, Y_{Km}^*)$ ;
6   end
7   Approximate  $G_{\hat{\nu}_\mu}(t)$  by  $M^{-1} \sum_{m=1}^M 1\{T_m^* \leq t\}$ ;
8 end

```

---

**3.2. Illustration**

To illustrate our proposed method, two examples are shown in Figure 1. In each, a meta-analysis is carried out on  $K = 3$  studies, and each study's variance  $\sigma_k^2$  generated from a inverse gamma distribution with a shape and scale parameter of 1. The data supplied from each of these hypothetical studies were generated from a normal distribution in which the true population mean was set at  $\mu = 0$  and the variance at  $\nu + \sigma_k^2$ . It is straightforward to construct plausibility functions for the individual studies, with data  $(Y_k, \sigma_k^2)$ , for  $k = 1, 2, 3$ , under the normality assumption (see, *e.g.*, Martin 2017), and these curves are plotted in gray in Figure 1. The black curve corresponds to the marginal plausibility function,  $\mu \mapsto \text{mpl}_y(\mu)$ , described in the previous section, evaluated using the Monte Carlo method outlined in Algorithm 1. Panels (a) and (b) correspond to  $\nu = 1$  and  $\nu = 2$ , respectively. Note that, as expected, the black curve is a “combination” of the three gray curves, with more influence coming from those gray curves that are tighter, corresponding to a more informative individual study.

**3.3. Theoretical properties**

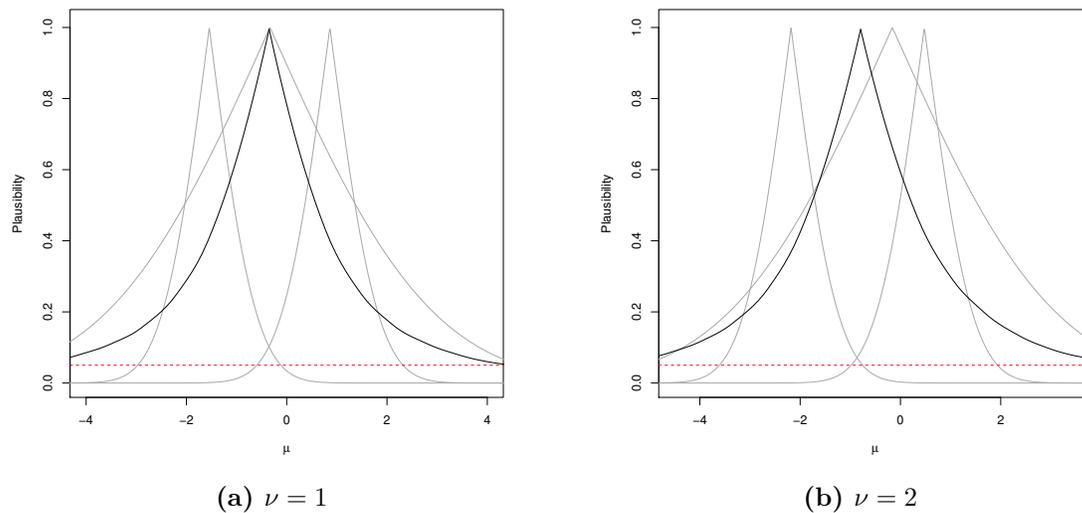
If the value of the nuisance parameter  $\nu$  were known, and used, in our construction of the (marginal) generalized inferential model for  $\mu$ , then the validity property, as stated in (3), would be immediate. For the practical case where  $\nu$  is unknown, we have recommended the inferential model with marginal point plausibility function (7), which involves a plug-in estimator. Our use, however, of this plug-in  $\hat{\nu}_\mu$ , complicates verification of the validity property. At the very least, under mild assumptions, our proposed generalized inferential model would be valid for large  $K$ , and the following theorem confirms this.

**Theorem 1:** Let  $Y^K = (Y_1, \dots, Y_K)$  be an independent sample from the random effects model,  $\mathbf{P}_{Y|\mu, \nu}$ , described above, where both  $\mu$  and  $\nu$  are unknown, but each within-study variance  $\sigma_k^2$  is known. Then the marginal plausibility function  $\text{mpl}_{Y^K}$  in (7) satisfies

$$\text{mpl}_{Y^K}(\mu) \rightarrow \text{Unif}(0, 1) \quad \text{in distribution under } \mathbf{P}_{Y|\mu, \nu} \text{ as } K \rightarrow \infty.$$

In particular, the marginal plausibility region  $\{\mu : \text{mpl}_{Y^K}(\mu) > \alpha\}$  has coverage probability approximately equal to  $1 - \alpha$ , for large  $K$ .

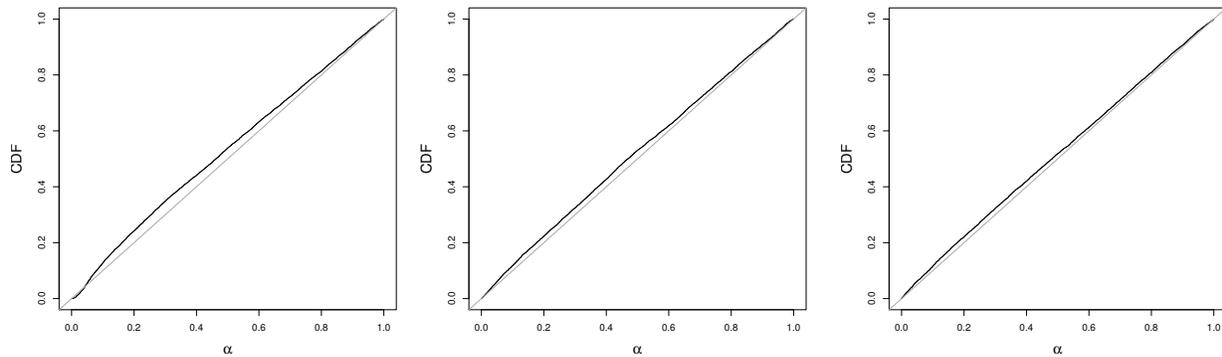
**Proof:** See the Appendix. □



**Figure 1: Examples of two simulated meta-analyses where the number of studies available  $K = 3$ . Plausibility functions associated with each individual study (in gray) and the combined plausibility function associated with our proposed inferential methods approach (in black). Marginal 95% plausibility intervals for  $\mu$  can be obtained where the combined plausibility intersects with  $\alpha = 0.05$ .**

The above theorem only says that the proposed solution is approximately valid for large  $K$ , but we do have reason to believe that the support for the proposed solution is actually stronger than this theorem suggests. Indeed, numerically, even for small  $K$ , the distribution of  $\text{mpl}_Y(\mu)$  is very close to uniform. As shown in Figure 2, based on 10,000 samples of the data pairs  $(Y_k, \sigma_k^2)$  from a small number of studies  $K = \{3, 4, 5\}$  and a high level of heterogeneity between studies  $\nu = 5$ , the distribution is close to uniform. There is some deviation to the left of uniform when the number of studies included for meta-analysis is particularly small,  $K = 3$ , but this is in the middle of the distribution, not in the lower tails (*e.g.*, around 0.05) where we would naturally be interested. Therefore, the method appears to be not only valid, but nearly exact.

It is natural to ask: why does our proposed method achieve this apparent higher-order of accuracy? At least intuitively, this can be answered by noticing that our proposed method has features in common with both the exact and higher-order asymptotically accurate methods described above. That is, by starting with the relative profile likelihood  $T_Y(\mu)$  in (6), we remove almost all of the dependence on the nuisance parameter; that is, by Wilks's theorem, the profile likelihood ratio has a known distribution—no nuisance parameter dependence—up to first order. This means that the exact distribution,  $G_\nu$ , of our  $T_Y(\mu)$  is roughly constant in  $\nu$ . Therefore, even though there is some remaining dependence on  $\nu$ , which is why a plug-in estimator is needed, it is not necessary that it be an especially accurate estimate. Ultimately, our final inferential model is built using the plug-in distribution  $G_{\hat{\nu}_\mu}$ , at each individual  $\mu$  value, which is very close to the exact distribution. It is this extra accuracy that leads to the superior practical performance in Figure 2 and Section 4, beyond what would be expected from the large- $K$  approximate validity result in Theorem 1.



**Figure 2:** Distribution function  $\alpha \mapsto \mathbb{P}_{Y|\mu,\nu}\{\text{mpl}_Y(\mu) \leq \alpha\}$  of  $\text{mpl}_Y(\mu)$  (black) compared with that of a  $\text{Unif}(0, 1)$  (gray) based on 10000 Monte Carlo samples from a normal-normal random effects model.. From left to right,  $K = 3, 4, 5$ .

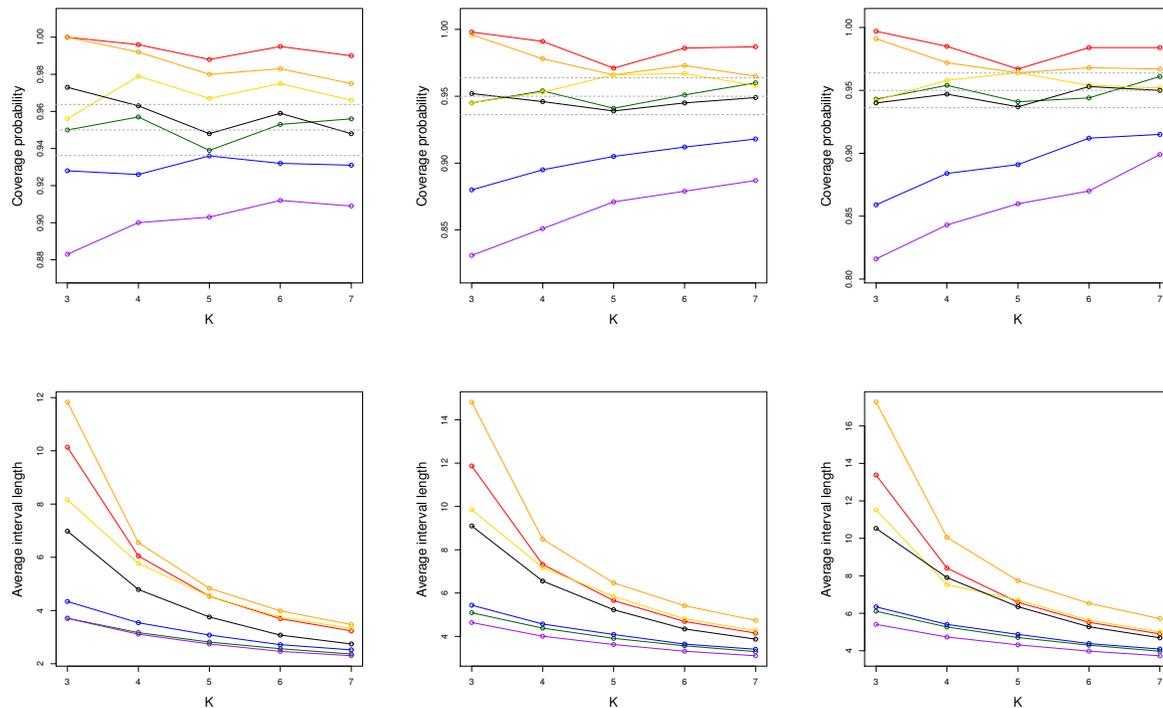
#### 4. Simulation Studies

Our simulations examine the performance of our proposed method, compared to that of other existing methods, in the case where  $K \leq 7$ . We generate our  $K$  study-level observations from the normal-normal random effects model, where the within-study variances—the  $\sigma_k^2$  values, for  $k = 1, \dots, K$ —are taken as fixed constants; following Gelman (2006), we generate these “fixed values” from an inverse gamma distribution with a shape and scale parameter of 1. Here we fix the overall effect at  $\mu = 0$  but vary the between-study variances as  $\nu \in \{1, 3, 5\}$ , so that we capture various settings from low to high levels of heterogeneity. For each combination of  $K$  and  $\nu$ , we repeat the experiment 1000 times to get estimates of the coverage probability and mean length of various 95% confidence intervals.

We compare the coverage properties of our approach (IM) against that of DerSimonian and Laird (1986, DL), the exact method in Michael *et al.* (2019, EX), the signed profile log likelihood ratio in Severini (2000, LK), its Skovgaard corrected cousin in Guolo (2012, SV), as implemented in the `metaLik` package in R (Guolo and Varin 2012), a traditional full Bayesian solution with a non-informative Jeffreys prior, as implemented in the `bayesmeta` package (Röver 2012), and, as a benchmark, an *oracle* procedure that knows the true value of  $\nu$  and uses the classical normal distribution theory for inference on  $\mu$ . For EX, we set the tuning parameter  $c_0$  to the values based on  $K$  as recommended in Michael *et al.* (2019).

As shown in Figure 3, our proposed generalized inferential method outperforms all the other methods—except, of course, the oracle—in terms of both coverage and mean interval length. The Bayes, exact and higher-order likelihood methods tend to have too high nominal coverage, and the others too low. The over-coverage seen here is consistent with the results shown in Michael *et al.* (2019). Oddly, when the between-study variance parameter is set to a higher value  $\nu = 5$ , so that the average heterogeneity among the 1000 simulations is high, the Skovgaard corrected signed profile log likelihood actually achieves nominal coverage across all small settings of  $K$ . This is in line with the results in Guolo (2012), in which the estimator is sensitive to the level of heterogeneity.

To further highlight this sensitivity of the higher-order likelihood approach, we ran our methods above in the same settings used in Guolo (2012); more specifically, we



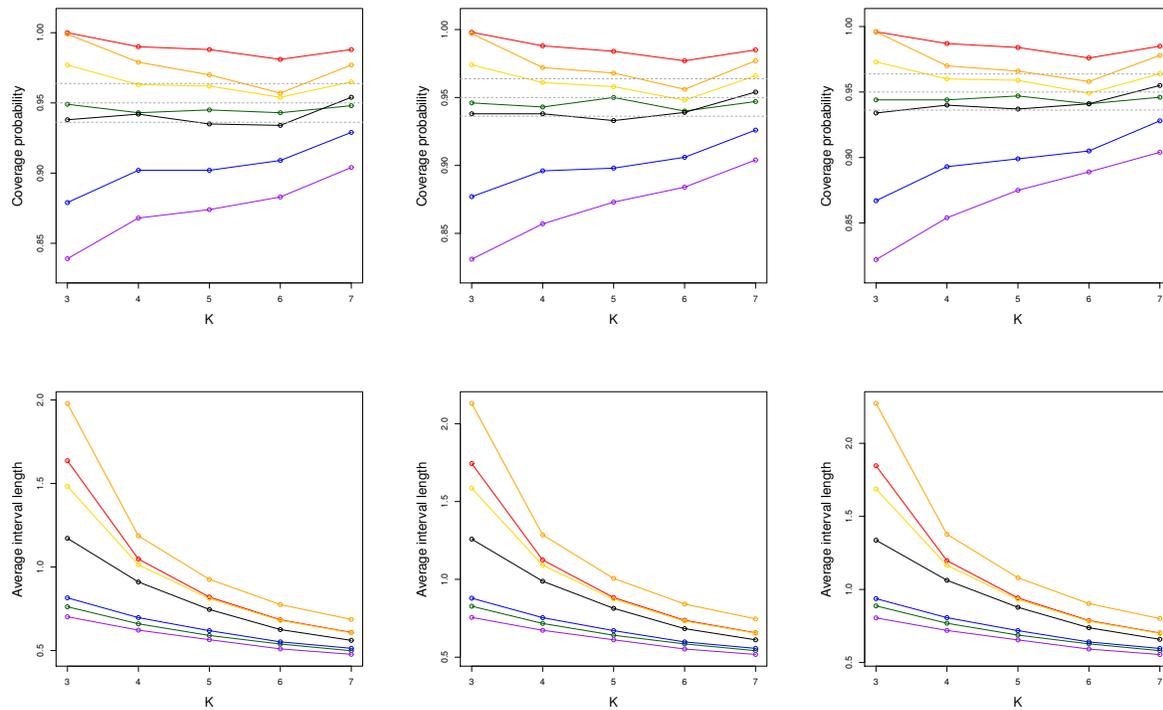
**Figure 3: Coverage probabilities and mean interval lengths across 15 different simulation settings for the number of studies available for meta-analysis  $K$  and the level of heterogeneity  $\nu$ . Results for DL (purple), LK (blue), oracle (green), SV (yellow), EX (orange), Bayes (red), and our proposed IM (black). From left to right, data are generated from a fixed between-study variance  $\nu \in \{1, 3, 5\}$ .**

re-examined the performance of our IM approach with  $K \in \{3, 4, 5, 6, 7\}$ ,  $\mu = 0.5$ , and  $\nu = \{0.08, 0.10, 0.12\}$ . We also generate the within-study variances for each study  $K$  from a uniform distribution on the interval 0.01 and 0.06 as done in Guolo (2012). As shown in Figure 4, our method still outperforms in these settings. Conversely, the under-coverage of DL and LK across these two different simulation settings in Figure 3 and Figure 4 are to be expected, as  $K$  is too small for the first-order asymptotic approximations to kick in. Moreover, our proposed method's strong coverage performance is not the result of having overly wide intervals: our mean lengths fall right in between those of the over- and under-coverage methods, and are quite close to that of the oracle as  $K$  becomes larger. Remarkably, these patterns hold across different heterogeneity levels as well.

## 5. Real Data Analyses

### 5.1. Changes in bone mineral density

To demonstrate how their inference procedure performs against popular meta-analytic techniques, Michael *et al.* (2019) carry out four separate meta-analyses on 59 randomised trials presented in Tai *et al.* (2015). These meta-analyses differ in two categories: (a) the specific bones from which the outcome measure, or the change in bone mineral density (BMD), was measured; and (b) the number studies included. The first meta-analysis consisted of



**Figure 4:** Coverage probabilities and mean interval lengths across 15 different simulation settings as done in Guolo (2012) for the number of studies available for meta-analysis  $K$  and the level of heterogeneity  $\nu$ . Results for DL (purple), LK (blue), oracle (green), SV (yellow), EX (orange), Bayes (red), and our proposed IM (black). From left to right, data are generated from a fixed between-study variance  $\nu \in \{0.08, 0.10, 0.12\}$  across various number of studies,  $K$ .

27 separate trials in which BMD changes were taken from the lumbar spine, followed by a meta-analysis of six trials of BMD changes from the hip, five from the forearm, and three from the total body. As shown in Table 1, for the first meta-analysis, our 95% plausibility interval almost matches the exact confidence interval from Michael *et al.* (2019), which also approximately aligns with that of the classical DerSimonian–Laird approach. This is no surprise as the number of studies  $K = 27$  itself is large. As for subsequent studies in which  $K \leq 6$ , the comparison between DerSimonian–Laird and the other two methods changes a lot and, in fact, sometimes leads to different scientific conclusions. For example, one would conclude a significant change in BMD from the forearm and total body meta-analyses based on DerSimonian–Laird, but conclude no significant change based on our method and that of Michael *et al.* (2019). Given that the latter two approaches have stronger theoretical support than the former, the difference in conclusions here might be indicative of the increased risk of false positives when using traditional meta-analytic techniques.

## 5.2. Risk of acute myocardial infarction

Here we consider a controversial example, one that called to question the use of meta-analyses in general (Egger and Smith 1995; Flather *et al.* 1997). Teo *et al.* (1991) conducted

**Table 1: Four meta-analyses on the effect of calcium supplements in changes in bone mineral density from Tai *et al.* (2015). Intervals based on three methods (DL, EX, and our IM) are reported, with interval lengths as subscripts.**

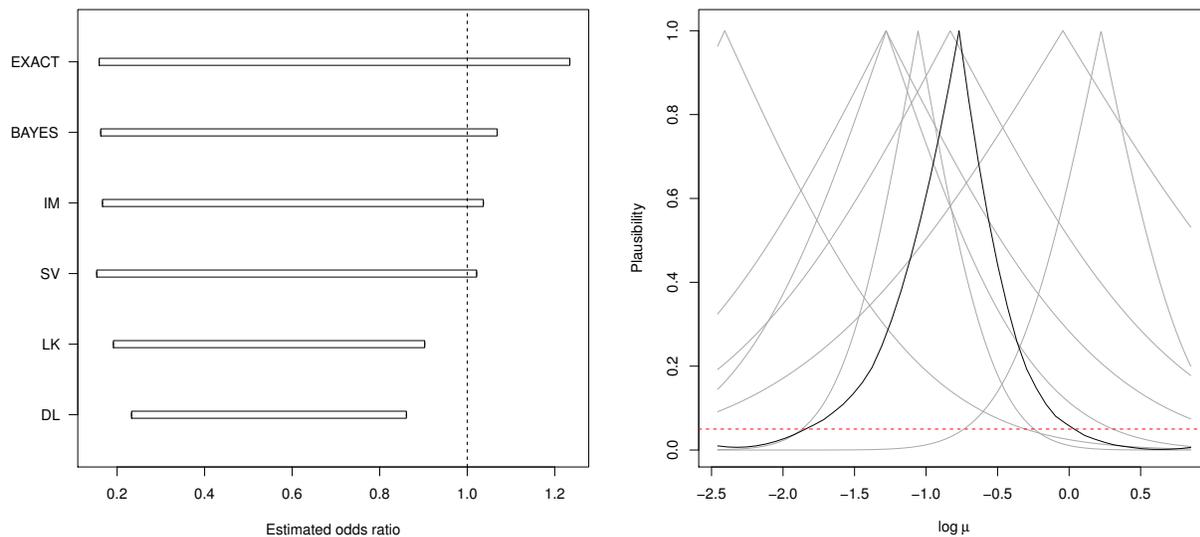
Study	$K$	DL	EX	IM
Lumbar spine	27	(0.828, 1.669) <sub>0.841</sub>	(0.768, 1.726) <sub>0.958</sub>	(0.811, 1.642) <sub>0.831</sub>
Total hip	6	(0.502, 1.847) <sub>1.345</sub>	(0.159, 2.246) <sub>2.087</sub>	(0.319, 2.131) <sub>1.812</sub>
Forearm	5	(0.209, 3.378) <sub>3.169</sub>	(−0.459, 4.124) <sub>4.583</sub>	(−0.426, 4.625) <sub>5.052</sub>
Total body	3	(0.268, 1.778) <sub>1.511</sub>	(−0.740, 2.796) <sub>3.536</sub>	(−0.486, 2.568) <sub>3.054</sub>

a meta-analysis of seven clinical trials that examined mortality across 1301 patients, 657 of which received intravenous magnesium therapy within 12 hours of hospitalization for acute myocardial infarction and 644 of which did not. In the original work, a fixed-effect method was used to combine the results from these seven randomized trials and arrive at a common odds ratio of 0.47, with 95% confidence interval (0.28, 0.79)—suggesting magnesium therapy to be highly effective in reducing mortality among this specific patient population. The expected drop in mortality, however, was refuted in a large-scale 58,050-patient follow-up study (Fourth International Infarct Survival Collaborative 1995) that estimated a common odds ratio of 1.06 with 95% confidence interval (1.00, 1.12). As a result, researchers raised concerns about meta-analytic techniques in general, citing issues around publication biases (Yusuf and Flather 1995) and high heterogeneity between studies (Flather *et al.* 1997). To address these problems, the canonical recommendation was to conduct sensitivity analyses via the use of multiple meta-analysis procedures, like that discussed below. Had such a precaution been taken, the fact that the Fourth International Infarct Survival Collaborative study lead to an alternative conclusion would not have been unforeseen.

It is also worth noting here that since the raw observations recorded in Teo *et al.* (1991) are in the form of a dichotomous outcome variable, we take the logarithm of the common odds ratio, between the mortality rate of patients that receive magnesium therapy and that of patients that do not, in order to conduct our meta-analysis as described in the competing procedures above. While there are other simplifications, *e.g.*, Van Houwelingen *et al.* (1993), we subscribe to the rationale in DerSimonian and Laird (1986) that regards the distribution of the log odds as approximately normal. Figure 5 thus compares the resulting interval estimates based on several meta-analytic procedures, namely, those assessed in Section 4. Note that the DerSimonian–Laird and signed profile log likelihood intervals approximate the original results from Teo *et al.* (1991). However, our proposed approach, along with the full Bayesian, the higher-order likelihood, and that in Michael *et al.* (2019), result in an odds ratio interval that suggests magnesium therapy does not significantly affect the short-term mortality of patients with acute myocardial infarction.

## 6. Conclusion

In this paper, we have considered an important and challenging problem, namely, valid statistical inference for meta-analyses that combine only a few studies. Again, the main obstacle is in dealing with the unknown between-study variance, in which there is only limited information in the few studies being combined. Our proposed solution is based on a recently



**Figure 5: (Left) Estimated odds ratio intervals for various meta-analytic techniques in the example described in Section 5.2. (Right) Combined plausibility function and respective IM interval on the log scale.**

proposed generalized inferential model framework, and we harness the power of profiling to construct a generalized association that is “almost” independent of the nuisance between-study variance. From there, we can use the exact distribution of the profile likelihood ratio, as the lack of sensitivity to the nuisance parameter means that it is not necessary to have an accurate plug-in estimator to achieve near-exact inference. In our numerical comparisons, we have demonstrated that the proposed inferential model solution outperforms existing methods in the literature, by being nearly exact and more efficient across a wide range of simulation settings, with few studies and both large and small between-study variance.

Given the strong performance in this meta-analysis application, it is natural to consider using the same generalized inferential model approach to solve other challenging problems. One that we have recently considered is when data come from a parametric model are corrupted by random censoring. The classical solution to this problem relies on the asymptotic normality of maximum likelihood estimators and, therefore, can only give approximately valid inference in an asymptotic sense. But the use of a likelihood ratio effectively marginalizes out the nuisance censoring distribution, so we end up in a position similar to that encountered in the present paper, the key difference being that the nuisance parameter is infinite-dimensional, which creates computational challenges. Preliminary results on this can be found in Cahoon and Martin (2019) and more details are forthcoming.

## Acknowledgments

The authors thank two anonymous reviewers for helpful feedback on an earlier version of the manuscript. They are also grateful to the editorial board of *Statistics and Applications* for the opportunity to make a contribution to this special issue in honor of the “Twin Statisticians,” Drs. Bikas and Bimal Sinha.

## References

- Balch, M. S., Martin, R. and Ferson, S. (2019). Satellite conjunction analysis and the false confidence theorem. *Proceedings of the Royal Society*, **A475(2227)**, 20180565.
- Barnard, G. A. (1995). Pivotal models and the fiducial argument. *International Statistical Review*, **63(3)**, 309–323.
- Cahoon, J. and Martin, R. (2019). Possibility measures for valid statistical inference based on censored data. In De Bock, J., de Campos, C.P., de Cooman, G., Quaeghebeur, E., and Wheeler, G. (Eds.), *Proceedings of the Eleventh International Symposium on Imprecise Probabilities: Theories and Applications*, Volume **103** of *Proceedings of Machine Learning Research*, pages 49–58, Thagaste, Ghent, Belgium. PMLR.
- Chung, Y., Rabe-Hesketh, S. and Choi, I. -H. (2013). Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in Medicine*, **32(23)**, 4071–4089.
- Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, **10(1)**, 101–129.
- Davey, J., Turner, R. M., Clarke, M. J. and Higgins, J. P. (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.
- Dawid, A. P. and Stone, M. (1982). The functional-model basis of fiducial inference. *The Annals of Statistics*, **10(4)**, 1054–1067.
- Dempster, A. P. (2008). The Dempster–Shafer calculus for statisticians. *International Journal of Approximate Reasoning*, **48(2)**, 365–377.
- DerSimonian, R. and Kacker, R. (2007). Random-effects model for meta-analysis of clinical trials: an update. *Contemporary Clinical Trials*, **28(2)**, 105–114.
- DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, **7(3)**, 177–188.
- Egger, M. and Smith, G. D. (1995). Misleading meta-analysis. *British Medical Journal Publishing Group*.
- Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. Hafner Publishing Co.
- Fletcher, M. D., Farkouh, M. E., Pogue, J. M. and Yusuf, S. (1997). Strengths and limitations of meta-analysis: larger studies may be more reliable. *Controlled Clinical Trials*, **18(6)**, 568–579.
- Follmann, D. A. and Proschan, M. A. (1999). Valid inference in random effects meta-analysis. *Biometrics*, **55(3)**, 732–737.
- Fourth International Infarct Survival Collaborative (1995). ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58,050 patients with suspected acute myocardial infarction. *Lancet*, **345(8951)**, 669–685.
- Fraser, D. A. S. (1968). *The Structure of Inference*. Wiley.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1(3)**, 515–534.
- Goodman, S. N. (1989). Meta-analysis and evidence. *Controlled Clinical Trials*, **10(2)**, 188–204.

- Guolo, A. (2012). Higher-order likelihood inference in meta-analysis and meta-regression. *Statistics in Medicine*, **31**(4), 313–327.
- Guolo, A. and Varin, C. (2012). The R package metaLik for likelihood inference in meta-analysis. *Journal of Statistical Software*, **50**(7), 1–14.
- Hardy, R. J. and Thompson, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine*, **15**(6), 619–629.
- Jackson, D., Bowden, J. and Baker, R. (2010). How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? *Journal of Statistical Planning and Inference*, **140**(4), 961–970.
- Liu, S., Tian, L., Lee, S. and Xie, Min-ge (2018). Exact inference on meta-analysis with generalized fixed-effects and random-effects models. *Biostatistics and Epidemiology*, **2**(1), 1–22.
- Martin, R. (2015). Plausibility functions and exact frequentist inference. *Journal of the American Statistical Association*, **110**(512), 1552–1561.
- Martin, R. (2017). A statistical inference course based on  $p$ -values. *The American Statistician*, **71**(2), 128–136.
- Martin, R. (2018). On an inferential model construction using generalized associations. *Journal of Statistical Planning and Inference*, **195**, 105–115.
- Martin, R. (2019). False confidence, non-additive beliefs, and valid statistical inference. *International Journal of Approximate Reasoning*, **113**, 39–73.
- Martin, R. and Liu, C. (2013). Inferential models: a framework for prior-free posterior probabilistic inference. *Journal of the American Statistical Association*, **108**(501), 301–313.
- Martin, R. and Liu, C. (2015a). Conditional inferential models: combining information for prior-free probabilistic inference. *Journal of the Royal Statistical Society*, **B77**(1), 195–217.
- Martin, R. and Liu, C. (2015b). Marginal inferential models: prior-free probabilistic inference on interest parameters. *Journal of the American Statistical Association*, **110**(512), 1621–1631.
- Martin, R. and Liu, C. (2016). *Inferential Models: Reasoning with Uncertainty*. CRC Press, Boca Raton, FL.
- Michael, H., Thornton, S., Xie, M. and Tian, L. (2019). Exact inference on the random-effects model for meta-analyses with few studies. *Biometrics*, **75**(2), 485–493.
- Paule, R. C. and Mandel, J. (1982). Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*, **87**(5), 377–385.
- Röver, C. (2017). Bayesian random-effects meta-analysis using the bayesmeta R package. <https://arxiv.org/abs/1711.08683/>.
- Severini, T. A. (2000). *Likelihood Methods in Statistics*. Oxford University Press.
- Severini, T. A. and Wong, W. H. (1992). Profile likelihood and conditionally parametric models. *The Annals of Statistics*, **20**(4), 1768–1802.
- Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.
- Shafer, G. (1987). Belief functions and possibility measures. In Bezdek, J., Editor, *The Analysis of Fuzzy Information*, Volume **1**, pages 51–84. CRC Press.

- Sidik, K. and Jonkman, J. N. (2007). A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine*, **26**(9), 1964–1981.
- Tai, V., Leung, W., Grey, A., Reid, I. R. and Bolland, M.J. (2015). Calcium intake and bone mineral density: systematic review and meta-analysis. *British Medical Journal*, **351**, 41–83.
- Taraldsen, G. and Lindqvist, B. H. (2013). Fiducial theory and optimal inference. *The Annals of Statistics*, **41**(1), 323–341.
- Teo, K. K., Yusuf, S., Collins, R., Held, P. H. and Peto, R. (1991). Effects of intravenous magnesium in suspected acute myocardial infarction: Overview of randomised trials. *British Medical Journal*, **303**(6816), 1499–1503.
- Tsui, K.-W. and Weerahandi, S. (1989). Generalized  $p$ -values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of the American Statistical Association*, **84**(406), 602–607.
- Van Houwelingen, H. C., Zwinderman, K. H. and Stijnen, T. (1993). A bivariate approach to meta-analysis. *Statistics in Medicine*, **12**(24), 2273–2284.
- Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D. and Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, **7**(1), 55–79.
- Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, **30**(3), 261–293.
- Wang, Y. and Tian, L. (2018). An efficient numerical algorithm for exact inference in meta analysis. *Journal of Statistical Computation and Simulation*, **88**(4), 646–656.
- Yusuf, S. and Flather, M. (1995). Magnesium in acute myocardial infarction. *British Medical Journal*, **310**, 751–752.

## APPENDIX

### Proof of Theorem 1

By definition of  $\text{mpl}_{YK}$  in (7), it is enough to show that  $G_{\hat{\nu}_\mu}^K(T_{YK}(\mu))$  converges in distribution to  $\text{Unif}(0, 1)$  under  $\mathbb{P}_{YK|\mu, \nu}$ ; note that here we insert the superscript “ $K$ ” to highlight the dependence on the number of studies,  $K$ . Then we can write  $G_{\hat{\nu}_\mu}^K(T_{YK}(\mu)) = G_\nu^K(T_{YK}(\mu)) + \Delta_K$ , where

$$\Delta_K = G_\nu^K(T_Y(\mu)) - G_{\hat{\nu}_\mu}^K(T_Y(\mu)),$$

with  $\hat{\nu}_\mu$ , the maximum likelihood estimate of the heterogeneity parameter at a fixed  $\mu$ , and  $\nu$  the true heterogeneity between studies. The key observation is that  $G_\nu^K(T_{YK}(\mu))$  is exactly uniformly distributed under  $\mathbb{P}_{YK|\mu, \nu}$ , so if we can show  $\Delta_K \rightarrow 0$  in  $\mathbb{P}_{YK|\mu, \nu}$ -probability, then the claim follows from Slutsky’s theorem.

Towards this, we clearly have

$$|\Delta_K| \leq \sup_{t \in [0, 1]} |G_\nu^K(t) - G_{\hat{\nu}_\mu}^K(t)|,$$

so we can prove our claim by showing the difference between the two distribution functions vanishes uniformly. But since these are distribution functions, it is enough to show that the difference vanishes pointwise, at each fixed  $t$ . Towards this, according to Guolo (2012), the meta-analysis problem is sufficiently regular that the classical first-order distribution theory applies; see, *e.g.*, Severini and Wong (1992, Sec. 4.6). In particular, this implies  $\hat{\nu}_\mu = \nu + O_P(K^{-1/2})$  which, in turn, implies that  $P_{Y^K|\mu,\nu}$  and  $P_{Y^K|\mu,\hat{\nu}_\mu}$  are mutually contiguous. Then the classical Wilks's theorem gives us

$$-2 \log T_{Y^K}(\mu) \rightarrow \text{ChiSq}(1) \quad \text{in distribution, as } K \rightarrow \infty, \quad (8)$$

under *both*  $P_{Y^K|\mu,\nu}$  and  $P_{Y^K|\mu,\hat{\nu}_\mu}$ . Therefore,

$$G_\nu^K(t) \rightarrow G^\infty(t) \quad \text{and} \quad G_{\hat{\nu}_\mu}^K(t) \rightarrow G^\infty(t),$$

where  $G^\infty$  is the limiting distribution function of  $T_{Y^K}(\mu)$  from (8). Now, if we write

$$|G_\nu^K(t) - G_{\hat{\nu}_\mu}^K(t)| \leq |G_\nu^K(t) - G^\infty(t)| + |G_{\hat{\nu}_\mu}^K(t) - G^\infty(t)|,$$

then we see the right-hand converges to 0 in  $P_{Y|\mu,\nu}$ -probability as  $K \rightarrow \infty$ . This implies  $\Delta_K \rightarrow 0$  from which  $G_{\hat{\nu}_\mu}^K(T_{Y^K}(\mu)) \rightarrow \text{Unif}(0,1)$  follows by Slutsky's theorem.