

# A REML method for the evidence-splitting model in network meta-analysis

Hans-Peter Piepho<sup>1</sup>  | Johannes Forkman<sup>2</sup>  | Waqas Ahmed Malik<sup>1</sup> 

<sup>1</sup>Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Stuttgart, Germany

<sup>2</sup>Department of Crop Production Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden

## Correspondence

Hans-Peter Piepho, Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Stuttgart, 70593, Germany. Email: [piepho@uni-hohenheim.de](mailto:piepho@uni-hohenheim.de)

## Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Number: PI 377/24-1

## Abstract

Checking for possible inconsistency between direct and indirect evidence is an important task in network meta-analysis. Recently, an evidence-splitting (ES) model has been proposed, that allows separating direct and indirect evidence in a network and hence assessing inconsistency. A salient feature of this model is that the variance for heterogeneity appears in both the mean and the variance structure. Thus, full maximum likelihood (ML) has been proposed for estimating the parameters of this model. Maximum likelihood is known to yield biased variance component estimates in linear mixed models, and this problem is expected to also affect the ES model. The purpose of the present paper, therefore, is to propose a method based on residual (or restricted) maximum likelihood (REML). Our simulation shows that this new method is quite competitive to methods based on full ML in terms of bias and mean squared error. In addition, some limitations of the ES model are discussed. While this model splits direct and indirect evidence, it is not a plausible model for the cause of inconsistency.

## KEYWORDS

inconsistency, maximum likelihood, mixed treatment comparisons, residual maximum likelihood

## 1 | INTRODUCTION

A key issue in network meta-analysis (NMA) is the potential occurrence of inconsistency between direct and indirect treatment comparisons. Several approaches have been proposed for detecting such direct–indirect evidence inconsistency as nicely reviewed by Shih and Tu.<sup>1</sup> These authors pointed out that previous proposals do not fully separate the direct and indirect evidence on a pairwise comparison of interest. They therefore proposed a new model, which they called the evidence-splitting (ES) model and which is based on the principle of independence between direct and indirect evidence. Splitting of direct and indirect evidence

is accomplished through inclusion of a regression term that depends on the variance for heterogeneity between trials (studies). The consequence of the presence of this variance parameter in both the mean and the variance structure of the model is that standard methods of estimation based on linear mixed models do not apply. The full maximum-likelihood (ML) method is a natural choice for estimation in this setting.<sup>1</sup> This method can be implemented in different ways, including structural equation modeling.<sup>2</sup> A well-known downside of ML estimation in linear mixed models is the bias of variance component estimates,<sup>3,4</sup> which can be substantial especially in arm-based network meta-analysis.<sup>5</sup> Hence, residual (or restricted) maximum likelihood

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Research Synthesis Methods* published by John Wiley & Sons Ltd.

(REML) is the method of choice for such models. However, as the variance is present in both the mean and the variance structure of the ES model, REML is not directly applicable.

Network meta-analysis is most commonly implemented using a contrast-based approach,<sup>6</sup> whereby in every trial each treatment arm is compared to a baseline treatment. The alternative is to use an arm-based approach. This approach amounts to fitting a two-way analysis-of-variance model with fixed effects for trials and treatments, and a random effect for the trial-by-treatment interaction.<sup>7</sup> The random effect represents heterogeneity between trials. Taking the trial main effect as fixed ensures that the principle of concurrent control is observed, that is, all inference on treatment effects is solely based on within-trial information, whereas between-trial information is not recovered.<sup>8,9</sup> The ES model proposed by Shih and Tu<sup>1</sup> is based on the contrast-based approach. Splitting the direct and indirect evidence involves linear combinations among all contrasts within a trial, ensuring that direct and indirect evidence are uncorrelated.<sup>1</sup> However, the authors also give an arm-based representation of the model, which is particularly convenient for network meta-analysis in case some trials involve a larger number of treatments. This is because the arm-based approach requires focusing only on the two treatment arms of interest, whereas the contrast-based approach requires forming linear combinations among all treatments per trial, for the purpose of evidence splitting.

The present paper proposes an iterative approximate method for estimating the ES model with both the contrast-based and arm-based representations, which allows REML estimation to be used. To assess the merit of this new method, we will compare it to ML-based estimation methods. The models and methods are laid down in Section 2. Two examples are considered in Section 3. In Section 4, we report on a small simulation study comparing the estimation methods. Some limitations of the ES model are pointed out in Section 5. The paper ends with some conclusions in Section 6.

## 2 | THE EVIDENCE-SPLITTING MODEL

Without loss of generality, we here give the models for a three-arm study involving treatments A, B, and C. This case is the simplest and relevant to the example considered in Section 3.1. For the purpose of exposition, the response is binomial, and empirical log-odds are analyzed under the assumption of approximate normality. However, the approach also applies for other response distributions and link functions, including the normal for continuous responses in combination with an identity

link, as considered in the example in Section 3.2. The justification we give for the ES model (in Appendices A and B) is somewhat different from that given by Shih and Tu,<sup>1</sup> who considered conditional distributions. We give our own compact derivation based on expected values here because this helps us focusing on the direct connections between the contrast-based and the arm-based form of the model. Furthermore, our derivation shows that the evidence can be split without using the ES model.

### 2.1 | The contrast-based basic model for NMA

Consider a network of trials with treatments A, B, and C. In trial  $i$ , at most three log-odds ratios can be observed:  $y_{iAB}$ ,  $y_{iAC}$ , and  $y_{iBC}$  for contrasts B versus A, C versus A, and C versus B, respectively. In a three-arm trial, the contrast-based model assumes that

$$\begin{pmatrix} y_{iAB} \\ y_{iAC} \end{pmatrix} \sim N \left[ \begin{pmatrix} d_{AB} \\ d_{AC} \end{pmatrix}, \begin{pmatrix} \tau^2 + v_{iA} + v_{iB} & \tau^2/2 + v_{iA} \\ \tau^2/2 + v_{iA} & \tau^2 + v_{iA} + v_{iC} \end{pmatrix} \right].$$

(Model 1. Contrast-based basic model for network meta-analysis), where  $d_{AB}$  and  $d_{AC}$  are the expected values of the two contrasts  $y_{iAB}$  and  $y_{iAC}$ ,  $\tau^2$  is the heterogeneity variance and  $v_{iA}$ ,  $v_{iB}$ ,  $v_{iC}$  are the outcome variances of A, B, and C, respectively. The variance for heterogeneity,  $\tau^2$ , is due to trial-by-treatment interaction. Note that the covariance among contrasts involves half the variance for heterogeneity,  $\tau^2/2$ , as well as the outcome variance that the two contrasts share. By comparison, the variance of each contrast involves the full variance for heterogeneity,  $\tau^2$ , as well as two outcome variances, since each contrast has contributions from two treatments. The model for the third contrast,  $y_{iBC}$ , can be written as  $y_{iBC} \sim N(d_{BC}, \tau^2 + v_{iB} + v_{iC})$ , where  $d_{BC} = d_{AC} - d_{AB}$  is the expected value. The outcome variances  $v_{iA}$ ,  $v_{iB}$ ,  $v_{iC}$  may differ between trials and also between treatments in the same trial, depending on the individual trial design and kind of summary statistic used as the response. They are known quantities obtained from the analyses of the individual trials. In a two-arm trial, just one of the three contrasts  $y_{iAB}$ ,  $y_{iAC}$ , and  $y_{iBC}$  is observed. This contrast is still normally distributed with expected value and variance according to Model 1.

### 2.2 | The contrast-based evidence-splitting model

Following Shih and Tu,<sup>1</sup> we consider evidence splitting for the C versus B contrast. For this comparison, the

contrast-based ES model for the  $i$ -th three-arm trial can be stated as

$$\begin{pmatrix} y_{iAB} \\ y_{iAC} \end{pmatrix} \sim N \left[ \begin{pmatrix} d_{AB} + x_{iAB}w \\ d_{AC} + x_{iAC}w \end{pmatrix}, \begin{pmatrix} \tau^2 + v_{iA} + v_{iB} & \tau^2/2 + v_{iA} \\ \tau^2/2 + v_{iA} & \tau^2 + v_{iA} + v_{iC} \end{pmatrix} \right].$$

(Model 2. Contrast-based ES model for network meta-analysis), and  $y_{iBC} \sim N(d_{BC} + w, \tau^2 + v_{iB} + v_{iC})$ , where  $d_{BC} = d_{AC} - d_{AB}$ . In this model,  $w$  is the discrepancy between the direct and the indirect evidence on the C versus B contrast. The latent covariates are given by

$$x_{iAB} = -\frac{\text{var}(y_{iAB}) - \text{cov}(y_{iAB}, y_{iAC})}{\text{var}(y_{iAB}) + \text{var}(y_{iAC}) - 2\text{cov}(y_{iAB}, y_{iAC})} \text{ and}$$

$$x_{iAC} = \frac{\text{var}(y_{iAC}) - \text{cov}(y_{iAB}, y_{iAC})}{\text{var}(y_{iAB}) + \text{var}(y_{iAC}) - 2\text{cov}(y_{iAB}, y_{iAC})}.$$

The role of the latent covariates  $x_{iAB}$  and  $x_{iAC}$  in Model 2 is to enable evidence splitting, that is, estimation of the effect  $w$ . We denote these covariates as latent because they depend on unknown variance parameters of the model. Note that the latent covariates are only needed for three-arm trials, but not for two-arm trials.<sup>1</sup> The variance components are the same as in Model 1. Specifically, the variance for heterogeneity,  $\tau^2$ , is unknown and must be estimated, whereas the outcome variances  $v_{iA}$ ,  $v_{iB}$ ,  $v_{iC}$  are known. Model 2 is readily extended to multi-arm trials with more than three treatments, but all treatments must be considered. For details, see Shih and Tu.<sup>1</sup> In Appendix A, we provide an alternative derivation of the contrast-based ES model and show that evidence splitting does not strictly require the ES model.

### 2.3 | The arm-based basic model for NMA

If  $y_{iA}$ ,  $y_{iB}$ ,  $y_{iC}$  denote the observed log-odds for treatments A, B, C in a three-arm trial  $i$ , then an arm-based model can be written as

$$\begin{pmatrix} y_{iA} \\ y_{iB} \\ y_{iC} \end{pmatrix} \sim N \left[ \begin{pmatrix} \mu + \beta_i + \gamma_A \\ \mu + \beta_i + \gamma_B \\ \mu + \beta_i + \gamma_C \end{pmatrix}, \begin{pmatrix} \tau^2/2 + v_{iA} & 0 & 0 \\ 0 & \tau^2/2 + v_{iB} & 0 \\ 0 & 0 & \tau^2/2 + v_{iC} \end{pmatrix} \right].$$

(Model 3. Arm-based basic model for network meta-analysis), where  $\mu$  is an overall intercept,  $\beta_i$  is the fixed main effect of the  $i$ -th trial, and  $\gamma_A$ ,  $\gamma_B$ ,  $\gamma_C$  are the fixed main effects for the treatments A, B, C. This representation of the arm-based model differs slightly from that given by Shih and Tu<sup>1</sup> in that we use a form that does not require

specifying a baseline treatment. It is of the form of a two-way analysis-of-variance model with factors treatment and trial and as such in our experience is very convenient for implementation in a linear mixed model package.<sup>7</sup> Note that even though the linear predictor is over-parameterized, and hence restrictions on estimates of the effects need to be imposed, the predictions are unique.

### 2.4 | The arm-based evidence-splitting model

We here focus on the B versus A contrast for better comparability with Shih and Tu.<sup>1</sup> For the  $i$ -th three-arm trial, the arm-based ES model can be written as<sup>1</sup>

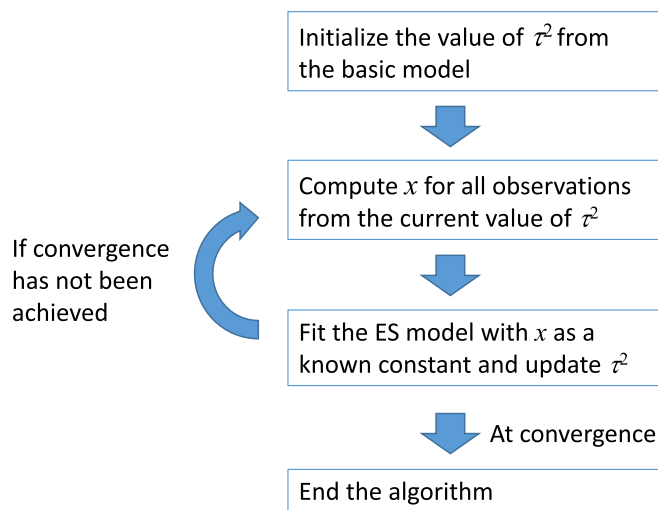
$$\begin{pmatrix} y_{iA} \\ y_{iB} \\ y_{iC} \end{pmatrix} \sim N \left[ \begin{pmatrix} \mu + \beta_i + \gamma_A + x_{iA}w \\ \mu + \beta_i + \gamma_B + x_{iB}w \\ \mu + \beta_i + \gamma_C \end{pmatrix}, \begin{pmatrix} \tau^2/2 + v_{iA} & 0 & 0 \\ 0 & \tau^2/2 + v_{iB} & 0 \\ 0 & 0 & \tau^2/2 + v_{iC} \end{pmatrix} \right].$$

(Model 4. Arm-based ES model for network meta-analysis), where  $w$  is the discrepancy between the direct and indirect evidence on the B versus A contrast and the latent covariates are defined as

$$x_{iA} = -\frac{\text{var}(y_{iA})}{\text{var}(y_{iA}) + \text{var}(y_{iB})} \text{ and}$$

$$x_{iB} = \frac{\text{var}(y_{iB})}{\text{var}(y_{iA}) + \text{var}(y_{iB})}.$$

Note that the latent covariate term is present only for the two treatments for which the evidence is split (A and B), but not on the third treatment (C). Also, the latent covariate is not needed for the two-arm trials. For trials with more than three treatments, the coding of the latent covariates is the same as shown here, regardless of the number of treatments in the trial.<sup>1</sup> Thus, whenever a trial has more than two treatments and comprises both treatments A and B, the latent covariates are added in the linear predictor for A and B in the form given above, otherwise the latent covariate is set to zero. This makes this representation of the model particularly convenient for meta-analyses involving trials with a larger number of treatments, which are very common in agriculture. In Appendix B, we provide a derivation of the arm-based ES model.



**FIGURE 1** Algorithm for obtaining iterated ML or REML estimate of the contrast-based ES Model 2 or the arm-based ES Model 4.

## 2.5 | Methods of estimation for the contrast- and arm-based models

The basic contrast-based Model 1 can be estimated by full ML and by REML. The ES model has the heterogeneity variance parameter  $\tau^2$  involved in both the latent covariate  $x$  and the variance structure. Thus, the model is no longer of the linear mixed form. As the parameter  $\tau^2$  is present both in the mean and the variance, standard REML is not an option, but full ML is. Shih and Tu<sup>1</sup> implemented this method using structural equation modeling. Here, we employ the NLMIXED procedure of SAS to implement the full ML method of Shih and Tu.<sup>1</sup> Our approach requires specifying the contrast-based Model 2 in the syntax of the procedure, based on which the full likelihood is then set up and maximized iteratively using adaptive Gaussian quadrature.<sup>10</sup>

An alternative is to employ an approximate iterative scheme using a linear mixed model in which the current value of the latent covariate  $x$  is considered a known constant. The latent covariate is computed for the current estimate of  $\tau^2$ , which is updated in the next iteration. These steps are repeated until the values of  $\tau^2$  in the latent covariate  $x$  and the current estimate based on the linear mixed model fit agree. If ML is used during iterations, we obtain an approximation of the full ML estimation. It is also possible to use REML in the iterations, and this is expected to be preferable due to a smaller bias in the estimation of  $\tau^2$ .<sup>3</sup> We implemented these iterative methods using the GLIMMIX procedure of SAS (see Supplemental Information). The iterative scheme is summarized in Figure 1.

The description above is focused on the contrast-based models. For the arm-based models, we have exactly

the same options as for the contrast-based models. In particular, the algorithm in Figure 1 can be applied in the same way, replacing the contrast-based Models 1 and 2 with the arm-based Models 3 and 4.

## 3 | EXAMPLES

### 3.1 | The Sclerotherapy NMA data

We consider the Sclerotherapy NMA data,<sup>11</sup> also used by Shih and Tu,<sup>1</sup> to illustrate their method. The trial network has two three-arm trials and 24 two-arm trials. The treatments were a control group (A), sclerotherapy (B), and beta-blocker (C). The focus of our analysis is on the direct-indirect evidence inconsistency for the C versus B contrast. The binomial outcome is the number  $r$  of participants suffering from bleeding. Trials with zero events or zero non-events in any of the treatment arms were handled by adding a value of 0.5 to the counts in both categories.<sup>12</sup> The outcome was transformed to log-odds ( $y_{iA}$ ,  $y_{iB}$ ,  $y_{iC}$ ) and analyzed assuming approximate normality. The outcome error variance for the log-odds of a treatment in a given trial was computed as  $v = 1/r + 1/(n-r)$ , where  $n$  is the binomial sample size.

Table 1 shows the estimates based on the basic Models 1 and 3. The REML-based analysis agrees between both models, whereas the ML-based analyses differ. The agreement of the REML estimates of  $\tau^2$  is expected, because the likelihood function being maximized is based on contrasts taking out all fixed effects in both cases.<sup>7</sup> It is noteworthy that the ML estimate of  $\tau^2$  is rather small for the arm-based model, indicating substantial downward bias (also see Section 4).

Table 2 shows the results for the ES models. The results for the contrast-based representation agree well with those in Shih and Tu<sup>1</sup> (see their Table 2) when full ML is used. Iterated ML provides a very good approximation. The arm-based and contrast-based results differ notably when full or iterated ML is used. By contrast, results agree between arm-based and contrast-based model when iterated REML is used. The heterogeneity variance estimate with ML methods is considerably smaller than with REML, again suggesting that ML methods are fraught with downward bias, particularly when an arm-based parameterization is used (see Section 4).

### 3.2 | A multi-environment variety trial

We here consider a series of 85 crop variety trials evaluating 13 varieties of spring wheat, conducted in Sweden

**TABLE 1** Parameter estimates (standard errors) for the basic models for network meta-analysis using different methods with the Sclerotherapy NMA data.<sup>1</sup>

	Contrast-based model (Model 1)		Arm-based model (Model 3)	
	ML	REML	ML	REML
Contrast $d_{BC}$	0.127 (0.430)	0.132 (0.451)	0.088 (0.277)	0.132 (0.451)
Heterogeneity variance $\tau^2$	0.877 (0.371)	0.999 (0.423)	0.192 (0.104)	0.999 (0.423)

**TABLE 2** Parameter estimates (standard errors) for ES models for C versus B contrast using different methods with the Sclerotherapy NMA data.<sup>1</sup>

	Contrast-based ES model (Model 2)				Arm-based ES model (Model 4)		
	Full ML (Shih and Tu) <sup>1</sup>	Full ML <sup>a</sup>	Iterated ML for $\tau^2$ <sup>b</sup>	Iterated REML for $\tau^2$ <sup>b</sup>	Full ML <sup>a</sup>	Iterated ML for $\tau^2$ <sup>b</sup>	Iterated REML for $\tau^2$ <sup>b</sup>
Direct estimate <sup>c</sup> $d_{BC}^{dir}$	-0.708 (0.787)	-0.709 (0.787)	-0.709 (0.787)	-0.729 (0.845)	-0.543 (0.519)	-0.543 (0.517)	-0.729 (0.845)
Indirect estimate <sup>c</sup> $d_{BC}^{ind}$	0.116 (0.505)	0.116 (0.505)	0.116 (0.506)	0.112 (0.543)	0.0958 (0.329)	0.0961 (0.329)	0.112 (0.543)
Inconsistency parameter $w$	-0.824 (0.935)	-0.825 (0.935)	-0.825 (0.936)	-0.841 (1.005)	-0.638 (0.616)	-0.640 (0.613)	-0.841 (1.005)
Heterogeneity variance $\tau^2$	0.847	0.847 (0.359)	0.853 (0.359)	1.037 (0.439)	0.192 (0.102)	0.194 (0.103)	1.037 (0.439)

<sup>a</sup>Using the NLMIXED procedure of SAS.<sup>b</sup>Using the GLIMMIX procedure of SAS.<sup>c</sup>For definitions see Appendices A and B.

during the five-year period 2002–2006 at multiple locations.<sup>13</sup> The purpose of these kinds of trials is to assess the value for cultivation and use (VCU) for each variety. The series of trials comprised one reference variety (with ID 9601). There was particular interest in comparing all candidate varieties to the reference variety. The dataset is rather unbalanced, with some varieties tested only in years 2005 and 2006. A salient feature of this dataset is that for some of the comparisons the mean of observed direct differences with the reference variety is quite different from the generalized least squares estimate of the same difference based on a suitable linear mixed model as will be described below. This apparent inconsistency between direct and indirect evidence hampered the communication of generalized squares estimates of variety means to growers, which led to the proposal of a new method to estimate the difference exclusively based on direct comparisons.<sup>13</sup> Here, we will investigate the magnitude and significance of the inconsistency.

Value for cultivation and use trials are routinely analyzed using only data of the current year using an unweighted two-stage approach.<sup>13,14</sup> In addition, a joint analysis across several years may be performed. We here consider both models, starting with the single-year analysis. We only use the arm-based approach, which is the

more convenient one due to the relatively large number of treatments. The available data consist of variety means per trial, each computed across four observations. In Swedish variety testing, information on the precision of the means is not carried forward. Thus, network meta-analysis for a single year was based on the model

$$y_{ij} = \mu + \gamma_j + \beta_i + e_{ij}$$

(Model 5. Arm-based basic model for year-wise analysis of series of crop variety trials), where  $y_{ij}$  is the mean of the  $j$ -th variety in the  $i$ -th trial and  $e_{ij}$  is a random residual comprising both trial-by-variety interaction (heterogeneity) and residual error associated with the mean. We assume normality for both the response and the residual. As the outcome error variances  $v_{ij}$  were not available, the residual was modeled as independent random normal deviate with constant variance  $\sigma_e^2$ .<sup>13</sup> Note that if the error variances  $v_{ij}$  were known, the residual would be modeled as  $\text{var}(e_{ij}) = \tau^2/2 + v_{ij}$ . Replacing this with  $\text{var}(e_{ij}) = \sigma_e^2$  implies the assumption that the error variances  $v_{ij}$  are constant. Assuming that the number of replications was the same in each trial, this assumption has some justification. While it is often found that there is some heterogeneity of error variance between variety trials, the effect of

**TABLE 3** Estimates (standard errors) of direct effects ( $d^{dir}$ ), indirect effects ( $d^{ind}$ ), and inconsistency ( $w$ ) in comparison of variety 20549 with reference variety 9601 and  $F$ -test for inconsistency in variety trial data.<sup>12</sup> Residual variance estimated by REML and ML.

Model Year(s)	Model 6				Model 8	
	2005		2006		2002–06	
	REML	ML	REML	ML	REML	ML
$\sigma_e^2$ Residual variance	90,762 (13,922)	67,673 (8964)	86,860 (12,537)	66,709 (8438)	74,472 (5475)	55,330 (3506)
<b>2005</b>						
$d^{ind}$	324 (537)	324 (464)			324 (487)	324 (420)
$d^{dir}$	−336 (426)	−336 (368)			−336 (386)	−336 (333)
$w$	−660 (686)	−660 (592)			−660 (621)	−660 (535)
<b>2006</b>						
$d^{ind}$			−1055 (332)	−1055 (291)	−1055 (308)	−1055 (265)
$d^{dir}$			−1044 (241)	−1044 (211)	−1044 (223)	−1044 (192)
$w$			−10 (410)	−10 (360)	−10 (380)	−10 (328)
<b>Test for inconsistency</b>						
$F$	0.93	1.24	0.00	0.00	0.56	0.76
$d.f.$	1	1	1	1	2	2
$p$ -value	0.3386	0.2682	0.9801	0.9773	0.5690	0.4684

accounting for such heterogeneity in a joint analysis over trials is often found to be relatively small.<sup>14</sup> Also, the constant variance assumption has a justification in randomization theory for multi-environment trials.<sup>15</sup>

As the residual has constant variance under the assumed model, the covariate  $x$  takes values  $-0.5$  and  $0.5$  for the two varieties to be compared when both are in the same trial, otherwise the value is zero. This is very simple to implement because  $x$  does not depend on the variance for heterogeneity here (i.e.,  $x$  is not latent) and analysis by REML is therefore exact. The extended model is

$$y_{ij} = \mu + \gamma_j + \beta_i + wx_{ij} + e_{ij}.$$

(Model 6. Arm-based ES model for year-wise analysis of series of crop variety trials), where  $w$  is the fixed effect for inconsistency.

Next, consider analysis across years. Locations varied between years, so we model trials as nested within years using<sup>13</sup>

$$y_{ijk} = \mu + \gamma_j + a_k + g_{jk} + \beta_{ik} + e_{ijk}$$

(Model 7. Arm-based basic model for analysis over years of series of crop variety trials), where  $y_{ijk}$  is the yield of the  $j$ -th variety in the  $k$ -th year and  $i$ -th trial,  $\mu$  is an overall intercept,  $\gamma_j$  is the main effect of the  $j$ -th variety,  $a_k$  is the main effect of the  $k$ -th year,  $g_{jk}$  is the variety-by-year interaction of the  $j$ -th variety with the  $k$ -th year,  $\beta_{ik}$  is the effect of the  $i$ -th trial nested within the  $k$ -th year, and  $e_{ijk}$  is a residual comprising both trial-by-variety interaction and error associated with  $y_{ijk}$ . The effects  $a_k$  and  $g_{jk}$  are modeled as random for routine analysis. If, however, we test for inconsistency within years, then all effects except the residual are fixed. In this case, we can make the replacement  $\gamma_{jk} = \gamma_j + a_k + g_{jk}$  and take this effect for the  $j$ -th variety and  $k$ -th year as fixed. With this parameterization, the joint model for simultaneously assessing inconsistency in the different years is

$$y_{ijk} = \mu + \gamma_{jk} + \beta_{ik} + w_k x_{ijk} + e_{ijk}$$

(Model 8: Arm-based ES model for analysis over years of series of crop variety trials), where  $w_k$  is the fixed effect for inconsistency in the  $k$ -th year. The covariate  $x_{ijk}$  takes values  $-0.5$  and  $0.5$  for the two varieties to be compared

when both are in the same trial, otherwise the value is zero.

To illustrate, consider the comparison of variety 20549 with the reference variety 9601. Variety 20549 was tested only in years 2005 and 2006, whereas the reference variety was tested in all years. Analyses were done both per year using Model 6 and across years using Model 8. The results are shown in Table 3. The direct difference estimates  $\hat{d}^{dir}$  (see Appendix B) agree exactly with those reported by Forkman<sup>13</sup> in his Table 2. The REML-based F-test for inconsistency is not significant, both for the individual year analyses ( $p = 0.3386$  on 1 d.f. for 2005,  $p = 0.9801$  on 1 d.f. for 2006) and the across-years analysis ( $p = 0.5690$  on 2 d.f.), giving no indication that the direct and indirect evidence for this comparison are inconsistent. In each of the two years, the direct and indirect difference estimates ( $\hat{d}^{dir}$  and  $\hat{d}^{ind}$ ; see Appendix B) are such that the corresponding generalized least squares estimates of the difference based on a model with fixed variety and trial effect ( $\hat{d} = -107$  in 2005,  $\hat{d} = -1053$  in 2006) fall between them as expected. In fact, since  $\hat{d}^{dir}$  and  $\hat{d}^{ind}$  based on the ES model are independent under normality, this generalized least squares estimate of the contrast,  $d$ , can be computed as the weighted average of  $\hat{d}^{dir}$  and  $\hat{d}^{ind}$ , with weights given by the inverses of the squared standard errors:

$$\hat{d} = \frac{[s.e.(\hat{d}^{dir})]^{-2}\hat{d}^{dir} + [s.e.(\hat{d}^{ind})]^{-2}\hat{d}^{ind}}{[s.e.(\hat{d}^{dir})]^{-2} + [s.e.(\hat{d}^{ind})]^{-2}} \text{ with}$$

$$s.e.(\hat{d}) = \left\{ [s.e.(\hat{d}^{dir})]^{-2} + [s.e.(\hat{d}^{ind})]^{-2} \right\}^{-1/2}.$$

Using this method,  $\hat{d} = -82$  (s.e. = 334) for the year 2005 and  $\hat{d} = -1048$  (s.e. = 195) for 2006. These estimates coincide exactly with the estimates using the standard NMA model. The standard errors, however, differ slightly between the two methods due to different residual variance estimates. Using the same residual variance estimate, the standard errors coincide exactly (not shown).

For comparison, we conducted the same analyses using full ML (Table 3). This yielded considerably smaller residual variance estimates, reflecting the large number of model degrees of freedom (123 in the across-years analysis) relative to the size of the dataset (498 observations). The bias can be calculated exactly in this case, observing that the REML estimator of the residual variance equals the residual mean  $MS_{residual}$  of an analysis of variance for the Model 6 or Model 8, hence is unbiased, and the ML estimator equals  $(n - p)n^{-1}MS_{residual}$ , where  $n$  is the sample size and  $p$  is

TABLE 4 Simulated bias and mean squared error (MSE, reported in brackets) of estimators for  $\tau^2$  and  $d_{BC}$  under consistent model using the Sclerotherapy NMA data<sup>1</sup> to parameterize the models (Models 1 and 3) for simulation, assuming  $\tau^2 = 0.3, 1.0$  and 3.0.

Method	Contrast-based consistent model (Model 1)		Arm-based consistent model (Model 3)	
	ML for $\tau^2$ <sup>a</sup>	REML for $\tau^2$ <sup>a</sup>	ML for $\tau^2$ <sup>a</sup>	REML for $\tau^2$ <sup>a</sup>
$\tau^2 = 0.3$				
$d_{BC}$	-0.0096 (0.0831)	-0.0080 (0.0831)	-0.0122 (0.0874)	-0.0088 (0.0831)
$\tau^2$ Heterogeneity variance	-0.1928 (0.0419)	-0.1712 (0.0350)	-0.2729 (0.0750)	-0.1683 (0.0338)
Convergence %	100	100	100	100
$\tau^2 = 1.0$				
$d_{BC}$	-0.0128 (0.1791)	-0.0118 (0.1799)	-0.0195 (0.1794)	-0.0117 (0.1799)
$\tau^2$ Heterogeneity variance	-0.6094 (0.3971)	-0.5599 (0.3440)	-0.8823 (0.7834)	-0.5598 (0.3439)
Convergence %	100	100	100	100
$\tau^2 = 3.0$				
$d_{BC}$	-0.0193 (0.4364)	-0.0187 (0.4345)	-0.0254 (0.4372)	-0.0187 (0.4345)
$\tau^2$ Heterogeneity variance	-1.8414 (3.5444)	-1.7189 (3.1358)	-2.5749 (6.6629)	-1.7189 (3.1358)
Convergence %	100	100	100	100

Note:

<sup>a</sup>Using the GLIMMIX procedure of SAS.

the number of fixed-effect parameters, meaning that the bias equals  $-n^{-1}p \times 100\%$  of  $MS_{residual}$ . For the across-years dataset we have  $n = 498$ ,  $p = 124$  such that  $p/n \approx 1/4$  and hence a bias of approximately  $-25\%$ . The ratio  $p/n$  is substantial yet mild by comparison with the usual setting in medical meta-analysis, where the number of treatments is often two in the majority of studies. For the Sclerotherapy NMA data in Section 3.1, we have  $p/n = 28/54 \approx 1/2$ . The REML-based estimator there is not a residual mean square, because the outcome variances  $v_{iA}$ ,  $v_{iB}$ ,  $v_{iC}$  enter the estimation procedure and the ES model is no longer of the linear mixed form, but the high  $p/n$  ratio still gives some rough indication of the order of magnitude of the discrepancy to be expected between iterated REML and ML estimates in an arm-based analysis. By the same token, bias in the contrast-based analysis is expected to be smaller, because trial effects are removed such that  $p$  equals only the number of treatment arms minus one and hence  $p/n$  is rather smaller for

**TABLE 5** Simulated bias and mean squared error (MSE, reported in brackets) of estimators for  $\tau^2$ ,  $d_{BC}^{dir}$ ,  $d_{BC}^{ind}$  and  $w$  under the ES model using the Sclerotherapy NMA data<sup>1</sup> to parameterize the model for simulation, assuming  $\tau^2 = 0.3, 1.0$  and  $3.0$ .

Method	Contrast-based ES model (Model 2)			Arm-based ES model (Model 4)		
	Full ML <sup>a</sup>	Iterated ML for $\tau^2$ <sup>b</sup>	Iterated REML for $\tau^2$ <sup>b</sup>	Full ML <sup>a</sup>	Iterated ML for $\tau^2$ <sup>b</sup>	Iterated REML for $\tau^2$ <sup>b</sup>
$\tau^2 = 0.3$						
$d_{BC}^{dir}$ Direct estimate	0.0143 (0.1191)	-0.0025 (0.3094)	-0.0032 (0.3118)	0.0422 (0.3149)	-0.0029 (0.3041)	-0.0032 (0.3118)
$d_{BC}^{ind}$ Indirect estimate	-0.0027 (0.3095)	0.0143 (0.1184)	0.0131 (0.1185)	-0.0520 (0.1237)	0.0150 (0.1259)	0.0131 (0.1180)
$w$ Inconsistency parameter	-0.0175 (0.4243)	-0.0168 (0.4239)	-0.0163 (0.4316)	0.0941 (0.3586)	-0.0150 (0.4275)	-0.0164 (0.4316)
$\tau^2$ Heterogeneity variance	-0.1008 (0.0276)	-0.1037 (0.0285)	-0.0406 (0.0256)	-0.2804 (0.0799)	-0.2806 (0.0801)	-0.0406 (0.0256)
Convergence %	96.40	94.61	98.70	99.87	37.85	97.77
$\tau^2 = 1.0$						
$d_{BC}^{dir}$ Direct estimate	0.0187 (0.2537)	-0.0007 (0.6746)	-0.0016 (0.6812)	0.0087 (0.6465)	0.0045 (0.6488)	-0.0016 (0.6812)
$d_{BC}^{ind}$ Indirect estimate	-0.0005 (0.6745)	0.0189 (0.2535)	0.0173 (0.2549)	0.0225 (0.2553)	0.0255 (0.2526)	0.0173 (0.2509)
$w$ Inconsistency parameter	-0.0195 (0.9347)	-0.0196 (0.9343)	-0.0188 (0.9431)	-0.0139 (0.8896)	-0.0210 (0.9013)	-0.0188 (0.9431)
$\tau^2$ Heterogeneity variance	-0.2625 (0.1675)	-0.2625 (0.1678)	-0.1148 (0.1413)	-0.7928 (0.6484)	-0.7936 (0.6498)	-0.1148 (0.1413)
Convergence %	99.95	100	100	100	95.50	99.98
$\tau^2 = 3.0$						
$d_{BC}^{dir}$ Direct estimate	0.0022 (1.6405)	0.0022 (1.6404)	0.0015 (1.6535)	0.0077 (1.5667)	0.0075 (1.5673)	0.0015 (1.6535)
$d_{BC}^{ind}$ Indirect estimate	0.0278 (0.6043)	0.0277 (0.6042)	0.0268 (0.6101)	0.0342 (0.5747)	0.0342 (0.5746)	0.0267 (0.6101)
$w$ Inconsistency parameter	-0.0256 (2.2596)	-0.0256 (2.2588)	-0.0253 (2.2785)	-0.0265 (2.1445)	-0.0267 (2.1444)	-0.0253 (2.2785)
$\tau^2$ Heterogeneity variance	-0.7949 (1.2206)	-0.7931 (1.2217)	-0.4263 (0.9456)	-2.2030 (4.9784)	-2.2033 (4.9798)	-0.4263 (0.9456)
Convergence %	100	100	100	100	100	100

<sup>a</sup>Using the NLMIXED procedure of SAS. This implements the full ML method of Shih and Tu.<sup>1</sup><sup>b</sup>Using the GLIMMIX procedure of SAS.

contrast-based analyses. This will be investigated further by simulation in the next section.

## 4 | SIMULATION

To investigate the bias and accuracy of alternative estimation methods for the ES model, we conducted a Monte Carlo simulation using the Sclerotherapy NMA data<sup>1</sup> to parameterize the arm-based model. The data were simulated under the null hypothesis of no inconsistency between direct and indirect evidence ( $w = 0$ ) for the C versus B contrast ( $d_{BC}$ ). Thus, assuming Model 3, we set

the fixed effects in the linear predictor equal to the estimates obtained by iterated REML. Random effects for heterogeneity were simulated from a normal distribution with variance  $\tau^2/2$ . The value of the parameter  $\tau^2$  was set to 0.3, 1.0 (close to the iterated REML estimate) and 3.0. The binomial response was generated using the binomial probability obtained by applying the inverse logit link to the linear predictor and using the sample size  $n$  as given in the example for the respective treatment and study. When non-convergence occurred, this was invariably because the estimator of  $\tau^2$  approached zero. In these cases, we replaced the analysis by one that fixed  $\tau^2$  at zero from the start. This way, convergence could be achieved



in all cases. The percentage of cases where convergence failed when  $\tau^2$  was a parameter to be estimated was recorded. The performance of the full ML method and the iterated ML and REML methods for the ES model was assessed in terms of the bias and mean squared error (MSE) of the estimators of  $\tau^2$ ,  $d_{BC}^{dir}$ ,  $d_{BC}^{ind}$ , and  $w$  (see Appendices A and B). Each of the three scenarios for  $\tau^2$  was assessed based on 10,000 simulation runs. The methods were assessed under both the contrast-based and the arm-based representation of the model for analysis. For comparison, we also investigated the analysis based on the consistent Models 1 and 3. Results are shown in Tables 4 and 5. Generally, for the ES model, the iterated REML method performs best in terms of both MSE and bias for  $\tau^2$  and effect estimates as well (Table 5). Overall, the performance of the iterated REML method is quite comparable to that of REML for analysis by the consistent model (Table 4). The method produces the same results for both contrast-based and arm-based approach, as is expected from corresponding results for the consistent model.<sup>5</sup> The key reason for this agreement between contrast-based and arm-based analysis lies in the fact that REML makes use of the residual contrasts, which remove any fixed effects, thus duly accounting for the degrees of freedom lost to fixed effects.<sup>3</sup> For the theoretical details we refer readers to Piepho et al.<sup>5</sup> The result carries over to our iterated REML method for the ES model, because on each iteration, a linear mixed model is fitted, conditioning on the current estimate of  $w$ . The ML-based methods suffer from larger bias for  $\tau^2$ , which also translates into larger MSE. The problem is most pronounced for the arm-based approach, which is due to the large number of fixed study effects. Generally, for both ML and REML methods, the MSE increased with the true value of  $\tau^2$ , which is expected by way of analogy with sample variances in simple random samples.<sup>3</sup> Interestingly, the point estimates of  $w$  have comparable bias and MSE for ML and REML methods. Note, however, that estimates of uncertainty for  $w$  estimates (model-based standard errors and confidence intervals) will be adversely affected by the inferior performance in estimating  $\tau^2$ . In summary, the results show that the iterated REML-approach is quite competitive. These results are restricted to smaller networks as investigated here. Performance in larger networks is certainly worth investigating in future.

## 5 | LIMITATIONS OF THE EVIDENCE-SPLITTING MODEL

The arm-based ES model has a single term  $wx_{ij}$  to detect the inconsistency between direct and indirect evidence.

The latent covariate  $x_{ij}$  takes non-zero values only in trials with direct evidence, so we will focus on such trials here. If we consider the direct evidence of the B vs. A contrast, the relevant latent covariate values for the  $i$ -th trial are  $x_{iA}$  and  $x_{iB}$ . The values of  $x_{iA}$  and  $x_{iB}$  depend on the variance components. While this model does lead to a clean separation of the direct and indirect evidence under the consistent two-way model, and as such is very useful, it does not represent a plausible data generating mechanism under inconsistency. To explain in more detail, consider the arm-based linear predictor under consistency, excluding the random effect for heterogeneity for clarity:

$$\eta_{ij} = \mu + \gamma_j + \beta_i$$

(Model 9. Arm-based linear predictor under consistency, excluding random effect for heterogeneity for clarity). This gives the expected value of the  $j$ -th treatment in the  $i$ -th trial on the linear-predictor scale under the consistent arm-based model. Adding the term for evidence splitting, the linear predictor becomes

$$\eta_{ij} = \mu + \gamma_j + \beta_i + wx_{ij}$$

(Model 10. Arm-based linear predictor for ES model, excluding random effect for heterogeneity for clarity). What this model would imply if taken at face value, is that in trials where both treatments A and B are present, the mean of treatment A is shifted by the amount  $wx_{iA}$  compared to trials not having both treatments, whereas simultaneously treatment B is shifted by the amount  $wx_{iB}$ . The parameter  $w$  itself equals the corresponding shift in the difference between the means of A and B. Thus, observing that  $x_{iB} - x_{iA} = 1$  under the ES model, the difference in the  $i$ -th trial, assumed to have both treatments A and B, is given by

$$\eta_{Bj} - \eta_{Aj} = \gamma_B - \gamma_A + w(x_{iB} - x_{iA}) = \gamma_B - \gamma_A + w.$$

The key point here is that  $x_{iB} - x_{iA} = 1$  regardless of the variance components. But under the ES model, the variance components additionally determine the exact values of  $x_{iA}$  and  $x_{iB}$ . Hence, the variance components dictate how large the shift of each of the two treatments is. Moreover, the shifts for A and B are further restricted to be opposite in sign, and also the shifts depend on the trial. While this is perfectly fine for the purpose of splitting the direct and indirect evidence under the consistent Model 9, it is too restrictive as a model explaining how any inconsistency may have come about in multi-arm trials, where treatments other than A and B do not experience a shift. Hence, the ES

model is best regarded as a convenient tool for tracing the separate flows of direct and indirect information for a particular contrast under the consistent Model 9, but not as a model describing a plausible data-generating mechanism under inconsistency.

As the focus of the analysis is on just the two treatments A and B in multi-arm trials where both are present, the most obvious choice for a plausible model under inconsistency is one that allows the direction and amounts of shift of A and B to be independent of any variance components. Clearly, the variance components cannot be causal agents in the genesis of inconsistency. The causal agents must be related to the conduct of a trial, its specific settings for the individual treatments and how this affects the treatment means. If there is an inconsistency regarding treatments A and B, then at least one of their effects must be shifted when they are in the same trial, as compared to the effect when they are not in the same trial. If that is a reasonable assumption, then it is also plausible to assume that both effects are shifted, and independently so, unless one has very specific evidence to show that only one of the two treatment effects would sustain a shift under inconsistency. In this scenario, one possible model under inconsistency may be formulated using the linear predictor

$$\eta_{ij} = \mu + \gamma_j + \beta_i + \delta_A z_{1ij} + \delta_B z_{2ij}.$$

(Model 11. An arm-based linear predictor with two parameters for inconsistency, excluding random effect for heterogeneity for clarity), where  $z_{1ij} = z_{2ij} = 0$  except for  $z_{1iA} = 1$  for the observation on treatment A and  $z_{2iB} = 1$  for the observation on treatment B if the  $i$ -th trial has both treatments A and B. The effect  $\delta_A$  is the shift of the effect for treatment A when in the same trial with B, and  $\delta_B$  is defined analogously. Apart from its realism, this Model 11 has two important advantages over the ES Model 10: (i) the covariates ( $z_{1ij}$  and  $z_{2ij}$ ) no longer depend on variance components, so the linear predictor is truly linear in the parameters and REML is directly available, and (ii) this model can also be fitted to individual-person (or individual-plot) data using a generalized linear mixed model (GLMM) for observed binomial counts, whereas the ES model only works for a linear mixed model (LMM) assuming approximate normality for the empirical logits. The same goes for other distributions and links with GLMM.

The one missing degree of freedom in the fixed effects for the ES Model 10 in comparison with Model 11 means that under this latter realistic scenario for inconsistency, the ES model cannot fully capture the inconsistency and hence the variance estimate for heterogeneity ( $\tau^2$ ) will sustain an upward bias. This, in turn, raises the question as to the best model to estimate  $\tau^2$ . The answer partly depends on the

purpose of the analysis. The models considered in this section provide three options: (i) the consistent Model 9, (ii) the ES Model 8 with one degree of freedom for inconsistency, and (iii) the extended Model 11 with two degrees of freedom for inconsistency. Using Model 9 provides a valid estimate of  $\tau^2$  under the null hypothesis of consistency. This estimate is fine if we assume that consistency holds and we just want a numerical check on the contributions of the direct and indirect evidence to the overall estimate of the B versus A contrast. This estimate is also fine for testing the null hypothesis of consistency, but it is not appropriate for obtaining valid standard errors of estimates of  $w$  under Model 10, which is why Shih and Tu proposed a ML estimation method based on Model 10.<sup>1</sup> In Section 2, we have continued along the same lines and developed alternative methods of estimation under the same Model 10, which we then compared by simulation under consistency in Section 4. The discussion in the present section suggests, however, that Model 10 may not be the best model to obtain a realistic estimate of  $\tau^2$  under inconsistency, because it only has one degree of freedom for inconsistency, whereas two degrees of freedom are required for a realistic representation of any inconsistency in relation to the two focal treatments. Hence, one may argue that Model 11 is preferable for estimating  $\tau^2$  if it is thought necessary to allow for inconsistency. Moreover, Model 11 may also be the better model for detecting inconsistency.

It should be re-iterated that the two-parameter Model 11 for inconsistency proposed here does not achieve the separation of direct and indirect evidence arising when the consistent Model 9 is assumed to hold, whereas this is achieved perfectly well with the ES Model 10. For example, if there are two three-arm trials, as in Example 1, then in addition to the direct evidence there is also indirect evidence on the treatment effects between the two three-arm trials due to the heterogeneity in the response variances  $v_{ij}$ . Model 11 cannot split that evidence, because the effects  $\delta_A$  and  $\delta_B$  simultaneously capture both sources of information, but the ES Model 10 can achieve the splitting via the single parameter  $w$ , which by construction captures the discrepancy between direct and indirect evidence. The ES Model 10 introduces the single additional parameter  $w$  as a device to trace and separate the flow of direct and indirect evidence on the B versus A contrast in fitting the consistent Model 9, without representing a plausible model for the genesis of any inconsistency. In all fairness, this is the main stated purpose of the ES model. By contrast, Model 11 introduces two parameters  $\delta_A$  and  $\delta_B$  in order to represent a realistic mechanism for inconsistency and simplify the estimation of the variance for heterogeneity. It may have better power for detecting inconsistency, but the model cannot

cleanly separate the direct and indirect evidence flowing in the consistent Model 9.

In this section, we have considered one specific model for inconsistency. There are, of course, many other models for inconsistency, and these are nicely reviewed in Shih and Tu.<sup>1</sup> Any of these models can be used as an alternative to Model 11 in analyses aimed at modeling inconsistency. Our purpose in introducing that model was not to advocate one specific model for inconsistency, but to illustrate the limitations of the ES model in representing a plausible mechanism for inconsistency.

In relation to option (i), it may be added that if consistency is assumed to hold, the split of the direct and indirect evidence could also be achieved as follows. First, fit the consistent Model 9 using all the data and obtain an estimate of the B versus A contrast and its associated variance. Denote these as  $\hat{d}_{AB}$  and  $\text{var}(\hat{d}_{AB})$ . Also, save the estimate of  $\tau^2$  from this analysis and denote this as  $\hat{\tau}_{all}^2$ . Second, only use the data on treatments A and B and only include trials having data on both A and B. Plug in the estimate  $\hat{\tau}_{all}^2$  and obtain the estimate of the B versus A contrast. This will be the estimate representing the direct evidence,  $\hat{d}_{AB}^{dir}$ , with associated variance  $\text{var}(\hat{d}_{AB}^{dir})$ . Third, noting that  $\hat{d}_{AB}$  must equal the weighted average of  $\hat{d}_{AB}^{dir}$  and  $\hat{d}_{AB}^{ind}$  with weights given by  $w_{dir} = \left[ \text{var}(\hat{d}_{AB}^{dir}) + \text{var}(\hat{d}_{AB}^{ind}) \right] / \text{var}(\hat{d}_{AB}^{dir})$  and  $w_{ind} = \left[ \text{var}(\hat{d}_{AB}^{dir}) + \text{var}(\hat{d}_{AB}^{ind}) \right] / \text{var}(\hat{d}_{AB}^{ind})$ , respectively, we can obtain  $\hat{d}_{AB}^{ind}$  and  $\text{var}(\hat{d}_{AB}^{ind})$  as follows:

$$\text{var}(\hat{d}_{AB}^{ind}) = \left( \frac{1}{\text{var}(\hat{d}_{AB})} - \frac{1}{\text{var}(\hat{d}_{AB}^{dir})} \right)^{-1}$$

and

$$\hat{d}_{AB}^{ind} = w_{ind}^{-1} \left( \hat{d}_{AB} - w_{dir} \hat{d}_{AB}^{dir} \right).$$

This reverse-engineering approach splits the evidence in exactly the same way as the ES model does when the same estimate of  $\tau^2$  is used. This equivalence makes it clear that the ES model is a tool for splitting the evidence under the consistent model.

## 6 | CONCLUSION

The ES model is useful for detection of inconsistency between direct and indirect evidence. As the model has

variance parameters in both the mean and the variance structure, full ML is the most natural approach to estimation. The downside of this method, as supported by our simulation, is substantial bias of the variance parameter estimator. Our proposed iterated REML estimator provides an efficient and less biased alternative. While the ES model is a convenient tool for comparing direct and indirect evidence, it does not provide a plausible model for the genesis of inconsistency. For this purpose, other models may be preferred.

## AUTHOR CONTRIBUTIONS

**Hans-Peter Piepho:** Conceptualization; methodology; investigation; funding acquisition; project administration; writing – original draft; writing – review and editing; formal analysis. **Johannes Forkman:** Methodology; investigation; formal analysis; visualization; writing – review and editing. **Waqas Ahmed Malik:** Software; formal analysis; writing – review and editing; investigation; validation.

## ACKNOWLEDGMENTS

This work was funded by grant PI 377/24-1 of the German Research Foundation (DFG). Open Access funding enabled and organized by Projekt DEAL.

## FUNDING INFORMATION

Hans-Peter Piepho and Waqas Malik were supported by the German Research Foundation (DFG) through grant PI 377/24-1.

## CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest.

## DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplemental material of this article.

## ORCID

Hans-Peter Piepho  <https://orcid.org/0000-0001-7813-2992>

Johannes Forkman  <https://orcid.org/0000-0002-5796-0710>

Waqas Ahmed Malik  <https://orcid.org/0000-0001-6455-5353>

## REFERENCES

- Shih MC, Tu YK. An evidence-splitting approach to evaluation of direct-indirect evidence inconsistency in network meta-analysis. *Res Synth Methods*. 2021;12:226-238.
- Tu YK, Wu YC. Using structural equation modelling for network meta-analysis. *BMC Med Res Methodol*. 2017;17:104.
- Searle SR, Casella G, McCulloch CE. *Variance components*. Wiley; 1992.

4. Jackson D, Law M, Stijnen T, Viechtbauer W, White IR. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Stat Med*. 2018;37(7):1059-1085.
5. Piepho HP, Madden LV, Roger J, Payne R, Williams ER. Estimating the variance for heterogeneity in arm-based network meta-analysis. *Pharm Stat*. 2018;17:264-277.
6. Lu G, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. *J Am Stat Assoc*. 2006;101:447-459.
7. Piepho HP, Williams ER, Madden LV. The use of two-way mixed models in multi-treatment meta-analysis. *Biometrics*. 2012;68:1269-1277.
8. Senn S. Hans van Houwelingen and the art of summing up. *Biom J*. 2010;52:85-94.
9. Piepho HP, Madden LV. How to observe the principle of concurrent control in an arm-based meta-analysis using SAS procedures GLIMMIX and BGLIMM. *Res Synth Methods*. 2022;13:821-828.
10. Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed effects model. *J Comput Graph Stat*. 1995;4:12-35.
11. Pagliaro L, D'Amico G, Sorensen TI, et al. Prevention of first bleeding in cirrhosis. A meta-analysis of randomized trials of nonsurgical treatment. *Ann Intern Med*. 1992;117:59-70.
12. Cox DR, Snell EJ. *Analysis of Binary Data*. Vol 32. Chapman and Hall/CRC Press; 1989.
13. Forkman J. The use of a reference variety for comparisons in incomplete series of crop variety trials. *J Appl Statistics*. 2013;40:2681-2698.
14. Möhring J, Piepho HP. Comparison of weighting in two-stage analyses of plant breeding trials. *Crop Sci*. 2009;49:1977-1988.
15. Caliński T, Czajka S, Kaczmarek Z, Krajewski P, Pilarczyk W. Analyzing multi-environment variety trials using randomization-derived mixed models. *Biometrics*. 2005;61:448-455.

## SUPPORTING INFORMATION

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

**How to cite this article:** Piepho H-P, Forkman J, Malik WA. A REML method for the evidence-splitting model in network meta-analysis. *Res Syn Meth*. 2024; 15(2):198-212. doi:[10.1002/jrsm.1679](https://doi.org/10.1002/jrsm.1679)

**APPENDIX A: DERIVATION OF THE CONTRAST-BASED ES MODEL**

In this section, we show that evidence splitting does not strictly require the ES model and also give an alternative derivation of the contrast-based ES model. The example that is considered in Section 3.1 uses treatment A as a baseline. Following Shih and Tu,<sup>1</sup> we focus on the contrast C versus B. Before moving to three-arm trials, consider the simpler case where all trials involve just two of three treatments. If the interest is in the C versus B contrast  $d_{BC}$ , then two-arm trials comprising treatments B and C (in short: BC-trials) provide direct evidence. Two-arm trials with treatments A and B (AB-trials) provide direct evidence on the B versus A contrast  $d_{AB}$ , and two-arm trials with treatments A and C (AC-trials) provide direct evidence on the C versus A contrast  $d_{AC}$ . In a consistent network, both AB- and AC-trials together provide indirect evidence on the C versus B contrast because  $d_{BC} = d_{AC} - d_{AB}$ . Hence, evidence splitting amounts to comparing the direct evidence provided by the BC-trials with the indirect evidence provided by the AB- and AC-trials.

Now assume that the network includes at least one three-arm trial (ABC-trial). Such trials comprise both direct and indirect evidence on the C versus B contrast. Hence, the evidence needs to be split within such trials. In a three-arm trial  $i$ , the direct evidence is given by

$$L_{i1} = y_{iAB} - y_{iAC} = -y_{iBC},$$

where  $y_{iBC}$  is the observed contrast (direct difference) between C and B. Now in the same three-arm trial we are looking for

$$L_{i2} = y_{iAB} + c_i y_{iAC}$$

with  $c_i$  chosen so that  $L_{i1}$  and  $L_{i2}$  are uncorrelated, and hence independent under normality, which we assume throughout. The reasoning here is that  $L_{i1}$  and  $L_{i2}$  provide a one-to-one transformation of  $y_{iAB}$  and  $y_{iAC}$ , representing the complete evidence provided by the  $i$ -th trial, and that since  $L_{i1}$  captures all direct evidence but no indirect evidence, then the uncorrelated  $L_{i2}$  must capture all the indirect evidence provided by the  $i$ -th trial.<sup>1</sup> Note that indirect evidence provided by  $L_{i2}$  comes from all the indirect comparisons between C and B that are available across the network of trials. Clearly, an indirect comparison always involves data from more than one trial, so  $L_{i2}$  in isolation is not informative about the C versus B contrast. For example, if the  $i$ -th trial provides a direct comparison of B versus A ( $y_{iAB}$ ) and another trial  $k$  provides a direct comparison of C versus A

( $y_{kAC}$ ), then the difference of these two direct comparisons ( $y_{iAB} - y_{kAC}$ ) provides an indirect comparison of B versus C. A large network of trials can provide a large number of such indirect comparisons. NMA seeks to exploit all of these indirect comparisons and combine them in an optimal way with the direct evidence.

Continuing our derivation, we find

$$\begin{aligned} \text{cov}(L_{i1}, L_{i2}) &= \text{var}(y_{iAB}) - c_i \text{var}(y_{iAC}) + (c_i - 1) \text{cov}(y_{iAB}, y_{iAC}) \\ &= c_i [\text{cov}(y_{iAB}, y_{iAC}) - \text{var}(y_{iAC})] \\ &\quad + [\text{var}(y_{iAB}) - \text{cov}(y_{iAB}, y_{iAC})] = 0 \\ \Rightarrow c_i &= \frac{\text{var}(y_{iAB}) - \text{cov}(y_{iAB}, y_{iAC})}{\text{var}(y_{iAC}) - \text{cov}(y_{iAB}, y_{iAC})} \end{aligned}$$

[Note that  $c_i$  is defined always because  $\text{var}(y_{iAC}) - \text{cov}(y_{iAB}, y_{iAC}) > 0$  always]. If there are  $n_{ABC}$  ABC-trials and  $n_{BC}$  BC-trials, then the direct evidence is given by the  $n_{ABC} + n_{BC}$  equations

$$L_{i1} = -d_{BC} + e_{i1} \quad (n_{ABC} \text{ equations for the ABC-trials})$$

$$y_{iBC} = d_{BC} + e_{iBC} \quad (n_{BC} \text{ equations for the BC-trials}).$$

where, according to Model 1,  $e_{i1}$  and  $e_{iBC}$  are both distributed as  $y_{iBC} \sim N(0, \tau^2 + v_{iB} + v_{iC})$ . From these equations, the contrast  $d_{BC}$  can be estimated using generalized least squares. This provides the direct-evidence estimate  $\hat{d}_{BC}^{dir}$ .

The indirect evidence is provided by the  $n_{ABC}$  ABC-trials, the  $n_{AB}$  AB-trials, and the  $n_{AC}$  AC-trials, through the  $n_{ABC} + n_{AB} + n_{AC}$  equations.

$$L_{i2} = d_{AB} + c_i d_{AC} + e_{i2} \quad (n_{ABC} \text{ equations for the ABC-trials})$$

$$y_{iAB} = d_{AB} + e_{iAB} \quad (n_{AB} \text{ equations for the AB-trials})$$

$$y_{iAC} = d_{AC} + e_{iAC} \quad (n_{AC} \text{ equations for the AC-trials})$$

where, by Model 1,  $e_{i2} \sim N[0, \tau^2 + v_{iA} + v_{iB} + c_i^2(\tau^2 + v_{iA} + v_{iC}) + 2c_i(\tau^2/2 + v_{iA})]$ ,  $e_{iAB} \sim N(0, \tau^2 + v_{iA} + v_{iB})$ , and  $e_{iAC} \sim N(0, \tau^2 + v_{iA} + v_{iC})$ . Let  $\hat{d}_{AB}^{ind}$  and  $\hat{d}_{AC}^{ind}$  denote the generalized least-squares estimates of  $d_{AB}$  and  $d_{AC}$ , respectively, based on these equations. The indirect-evidence estimate  $\hat{d}_{BC}^{ind}$  is computed as  $\hat{d}_{BC}^{ind} = \hat{d}_{AC}^{ind} - \hat{d}_{AB}^{ind}$ .

Since the direct- and indirect-evidence estimates  $\hat{d}_{BC}^{dir}$  and  $\hat{d}_{BC}^{ind}$  are uncorrelated, the generalized least-squares estimate of  $d_{BC}$  in Model 1 is the weighted average of  $\hat{d}_{BC}^{dir}$  and  $\hat{d}_{BC}^{ind}$ , using their inverse variances as weights. We have thus achieved the split of the evidence for  $d_{BC}$  in Model 1.

By Model 1, the direct- and indirect estimators are consistent, that is,  $E(\widehat{d}_{BC}^{dir}) = E(\widehat{d}_{BC}^{ind})$ . As we have seen, the ES model by Shih and Tu is not needed for splitting the evidence. However, their model allows for inconsistency between the direct and indirect evidence.<sup>1</sup> This is accomplished by the assumption that  $E(\widehat{d}_{BC}^{dir}) = E(\widehat{d}_{BC}^{ind}) + w$ , where  $w$  is the discrepancy between the direct and indirect evidence for the C versus B contrast. Under the null hypothesis of consistency between the direct and indirect evidence, we have  $w = 0$ . We shall now provide a simple justification for the ES model.

Consider an ABC-trial ( $i = 1$ ) and an AC-trial ( $i = 2$ ). Using the ABC-trial,  $\widehat{d}_{BC}^{dir} = y_{1AC} - y_{1AB}$ . Since  $\widehat{d}_{AC}^{ind} = y_{2AC}$ , we have  $\widehat{d}_{AB}^{ind} = L_{12} - c_1 \widehat{d}_{AC}^{ind}$  and

$$\begin{aligned}\widehat{d}_{BC}^{ind} &= \widehat{d}_{AC}^{ind} - \widehat{d}_{AB}^{ind} = y_{2AC} - y_{1AB} - c_1 y_{1AC} + c_1 y_{2AC} \\ &= (y_{1AC} - y_{1AB}) + (y_{2AC} - y_{1AC} + c_1 y_{2AC} - c_1 y_{1AC}) \\ &= \widehat{d}_{BC}^{dir} + (1 + c_1)(y_{2AC} - y_{1AC}).\end{aligned}$$

Now taking expectations, using  $w = E(\widehat{d}_{BC}^{dir}) - E(\widehat{d}_{BC}^{ind})$ , we find  $w = -(1 + c_1)E(y_{2AC} - y_{1AC})$ . Thus  $E(y_{iAC}) = d_{AC} + \frac{1}{1+c_1}w$ , where  $d_{AC} = E(y_{2AC})$ . As a result, the expected value for the contrast between C and A is  $(1 + c_1)^{-1}w$  units larger in an ABC-trial than in an AC-trial.

Similarly, we may consider the ABC-trial ( $i = 1$ ) and an AB-trial ( $i = 3$ ). Since  $\widehat{d}_{AB}^{ind} = y_{3AB}$  and  $\widehat{d}_{AC}^{ind} = (L_{12} - \widehat{d}_{AB}^{ind})/c_1$ , we have

$$\begin{aligned}\widehat{d}_{BC}^{ind} &= \widehat{d}_{AC}^{ind} - \widehat{d}_{AB}^{ind} = \frac{1}{c_1}(y_{1AB} + c_1 y_{1AC}) - \left(\frac{1}{c_1} + 1\right)y_{3AB} \\ &= (y_{1AC} - y_{1AB}) + \frac{c_1 + 1}{c_1}y_{1AB} - \frac{c_1 + 1}{c_1}y_{3AB} \\ &= \widehat{d}_{BC}^{dir} - \frac{c_1 + 1}{c_1}(y_{3AB} - y_{1AB})\end{aligned}$$

The expected difference between the direct- and indirect-evidence estimates is  $w = \frac{1+c_1}{c_1}E(y_{3AB} - y_{1AB})$ , which gives  $E(y_{1AB}) = d_{AB} - \frac{c_1}{1+c_1}w$ . In conclusion, the expected value for the contrast between B and C is  $c_i/(1 + c_i)$  smaller in an ABC-trial than in an AB-trial. Hence,

$$E\begin{pmatrix} y_{iAB} \\ y_{iAC} \end{pmatrix} = \begin{pmatrix} d_{AB} + x_{iAB}w \\ d_{AC} + x_{iAC}w \end{pmatrix}$$

with latent covariates

$$x_{iAB} = -\frac{\text{var}(y_{iAB}) - \text{cov}(y_{iAB}, y_{iAC})}{\text{var}(y_{iAB}) + \text{var}(y_{iAC}) - 2\text{cov}(y_{iAB}, y_{iAC})} \quad \text{and}$$

$$x_{iAC} = \frac{\text{var}(y_{iAC}) - \text{cov}(y_{iAB}, y_{iAC})}{\text{var}(y_{iAB}) + \text{var}(y_{iAC}) - 2\text{cov}(y_{iAB}, y_{iAC})},$$

where the variances and covariances are the same as given in Model 1. These definitions of the latent covariates are needed only for the three-arm trials in the example in Section 3.1. In all other studies, involving only two treatments, the covariate  $x$  equals 0, except when the two treatments are B and C, in which case  $x$  equals 1. Thus, the contrast-based ES model is different from Model 1, since a latent covariate  $x$  is added to assess the inconsistency between direct and indirect evidence for the C versus B contrast. Hence, the contrast-based ES model for the  $i$ -th three-arm trial can be stated as in Model 2. The variance components are the same as in Model 1. Specifically, the variance for heterogeneity,  $\tau^2$ , is unknown and must be estimated, whereas the outcome variances  $v_{iA}$ ,  $v_{iB}$ ,  $v_{iC}$  are known. Model 2 is readily extended to multi-arm trials with more than three treatments, but all treatments must be considered. For details, see Shih and Tu.<sup>1</sup>

## APPENDIX B: DERIVATION OF THE ARM-BASED ES MODEL

Here, we provide a derivation of the arm-based ES model. Assume we want to split the evidence for the comparison of treatments A and B. Here, we focus on this contrast, rather than on the C versus B contrast considered in the example in Section 3.1, for better comparability with the derivations given by Shih and Tu.<sup>1</sup> In a multi-arm trial comprising both A and B, we consider two linear combinations, that is,

$$M_{i1} = y_{iB} - y_{iA} \quad \text{and}$$

$$M_{i2} = y_{iB} + f_i y_{iA}$$

with  $f_i$  chosen so that  $\text{cov}(M_{i1}, M_{i2}) = 0$ .  $M_{i1}$  contains the direct evidence and  $M_{i2}$  the indirect evidence on the B versus A contrast. Under the assumption of independence between  $y_{iA}$  and  $y_{iB}$

$$\text{cov}(M_{i1}, M_{i2}) = \text{var}(y_{iB}) - f_i \text{var}(y_{iA}).$$

Thus, to ensure  $\text{cov}(M_{i1}, M_{i2}) = 0$  we require

$$f_i = \frac{\text{var}(y_{iB})}{\text{var}(y_{iA})}.$$

With this choice of  $f_i$ ,  $M_{i1}$  is uncorrelated with  $M_{i2}$ , the observed responses of treatment C in the  $i$ -th trial, and data from any other trials. Note that one could separately analyze  $M_{i1}$  from all three-arm trials together with all AB-trials to extract all direct evidence on the B versus A contrast. Likewise,  $M_{i2}$  and data on C from all three-arm trials together with the data from all AC- and BC-trials could be analyzed separately to extract all indirect evidence. However, a joint analysis is needed to assess the inconsistency between both estimates. To develop a model for this purpose, we now consider the expected values for  $M_{i1}$  and  $M_{i2}$ . First, we express the expectations in terms of the contrast between A and B and, as with the contrast-based approach, defining  $d_{AB}^{dir}$  to be the expected contrast based on direct evidence and  $d_{AB}^{ind}$  the expected contrast based on the indirect evidence. Again, under consistency  $d_{AB}^{dir} = d_{AB}^{ind}$ , whereas under inconsistency  $d_{AB}^{dir} \neq d_{AB}^{ind}$ . As in the contrast-based ES model (Section 2.2), we use the parameterization  $d_{AB}^{dir} = d_{AB}^{ind} + w$ , thus setting  $d_{AB}^{dir} = \gamma_B - \gamma_A + w$  and  $d_{AB}^{ind} = \gamma_B - \gamma_A$ , where  $\gamma_A$  and  $\gamma_B$  are the effects of treatments A and B (see Section 2.3). Hence, the expectations for  $M_{i1}$  and  $M_{i2}$  are (ignoring intercept  $\mu$  and trial effects  $\beta_i$  for simplicity)

$$E \begin{pmatrix} M_{i1} \\ M_{i2} \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ f_i & 1 \end{pmatrix} \times E \begin{pmatrix} y_{iA} \\ y_{iB} \end{pmatrix} = \begin{pmatrix} \gamma_B - \gamma_A + w \\ f_i \gamma_A + \gamma_B \end{pmatrix}.$$

Note that for the indirect evidence  $M_{i2}$  we are using the definitions  $E(y_{iA}) = \gamma_A$  and  $E(y_{iB}) = \gamma_B$  (again ignoring intercept and trial effects). By comparison, we cannot use these definitions in the model for the direct evidence  $M_{i1}$ , because this also involves  $w$ . This specification is analogous to that for the contrast-based ES model in Section 2.2. From this model for the expected values of  $M_{i1}$  and  $M_{i2}$ , we may obtain the expected values for the observed treatment responses as

$$E \begin{pmatrix} y_{iA} \\ y_{iB} \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ f_i & 1 \end{pmatrix}^{-1} \begin{pmatrix} \gamma_B - \gamma_A + w \\ f_i \gamma_A + \gamma_B \end{pmatrix}.$$

It can be shown that

$$\begin{pmatrix} -1 & 1 \\ f_i & 1 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{f_i + 1} & \frac{1}{f_i + 1} \\ \frac{f_i}{f_i + 1} & \frac{1}{f_i + 1} \end{pmatrix}.$$

Hence

$$\begin{aligned} E \begin{pmatrix} y_{iA} \\ y_{iB} \end{pmatrix} &= \begin{pmatrix} \frac{1}{f_i + 1} & \frac{1}{f_i + 1} \\ \frac{f_i}{f_i + 1} & \frac{1}{f_i + 1} \end{pmatrix} \begin{pmatrix} \gamma_B - \gamma_A + w \\ f_i \gamma_A + \gamma_B \end{pmatrix} \\ &= \begin{pmatrix} \gamma_A - \frac{1}{f_i + 1} w \\ \gamma_B + \frac{f_i}{f_i + 1} w \end{pmatrix} = \begin{pmatrix} \gamma_A + x_{iA} w \\ \gamma_B + x_{iB} w \end{pmatrix} \end{aligned}$$

with

$$x_{iA} = -\frac{\text{var}(y_{iA})}{\text{var}(y_{iA}) + \text{var}(y_{iB})} \text{ and}$$

$$x_{iB} = \frac{\text{var}(y_{iB})}{\text{var}(y_{iA}) + \text{var}(y_{iB})},$$

where the variances are given in Model 3. The latent covariate  $x$  is set to zero in all trials comprising only one of the two treatments A and B, and also for treatments other than A and B in the multi-arm trials involving both A and B, because the corresponding observations are not involved in the evidence splitting. The treatment effects given above for  $y_{iA}$  and  $y_{iB}$ , that is,  $\gamma_A + x_{iA} w$  and  $\gamma_B + x_{iB} w$ , are also given by Shih and Tu.<sup>1</sup> These effects differ from those in Model 3, as a latent covariate  $x$  has been added to assess any inconsistency between direct and indirect evidence for the B versus A contrast. The full linear predictor for the expectation of the  $j$ -th treatment in the  $i$ -th trial is  $\mu + \gamma_j + \beta_i + x_{ij} w$ . Hence, for the  $i$ -th three-arm trial, the arm-based ES model can be written as in Model 4.