Title: Assessing the consistency assumptions underlying network meta-regression using aggregate data.

Short title: Assessing consistency in network meta-regression.

Article type: Research article.

Authors and affiliations: Sarah Donegan[1], Sofia Dias[2], Nicky J. Welton[2].

[1]Department of Biostatistics, Waterhouse Building, University of Liverpool, 1-5 Brownlow Street, Liverpool, L69 3GL, UK.

[2]School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol, BS8 2PS, UK.

Corresponding author details:

Sarah Donegan

Department of Biostatistics, Waterhouse Building, University of Liverpool, 1-5 Brownlow Street, Liverpool, L69 3GL, UK.

Email: sarah.donegan@liverpool.ac.uk

Tel: +44 151 706 4277

Fax: +44 151 282 4721

Contributions of authors: SDo proposed extending the existing node-splitting models proposed by SDi and NW and inconsistency models to include treatment by covariate interactions. NW proposed additional modelling extensions. SDo carried out the analysis and wrote the first draft of the manuscript. SDi and NW provided statistical guidance and commented on the manuscript.

Word count: 6,182

Number of Figures: 5

Number of colour figures: 2

Number of tables: 7

Number of supplementary Figures: 0

Number of supplementary tables: 8

Funding

This research was supported by the Medical Research Council (grant number MR/K021435/1) as part of a career development award in biostatistics awarded to SDo.

Conflicts of interest: The authors have declared no competing interests exist.

Abstract

Word count: 250

When numerous treatments exist for a disease (treatments *1*, *2*, *3* etc.), network meta-regression (NMR) examines whether each relative treatment effect (e.g. mean difference for *2* vs. *1*, *3* vs. *1*, *3* vs. *2* etc.) differs according to a covariate (e.g. disease severity). Two consistency assumptions underlie NMR: consistency of the treatment effects at the covariate value zero and consistency of the regression coefficients for the treatment by covariate interaction. The NMR results may be unreliable when the assumptions do not hold. Furthermore, interactions may exist but are not found because inconsistency of the coefficients is masking them; for example, when the treatment effect increases as the covariate increases using direct evidence but the effect decreases with the increasing covariate using indirect evidence.

We outline existing NMR models that incorporate different types of treatment by covariate interaction. We then introduce models that can be used to assess the consistency assumptions underlying NMR for aggregate data. We extend existing node-splitting models, the unrelated mean effects inconsistency model and the design by treatment inconsistency model to incorporate covariate interactions. We propose models for assessing both consistency assumptions simultaneously and models for assessing each of the assumptions in turn to gain a more thorough understanding of consistency.

We apply the methods in a Bayesian framework to trial-level data comparing anti-malarial treatments using the covariate *average age*, and to four fabricated datasets to demonstrate key scenarios.

We discuss the pros and cons of the methods and important considerations when applying models to aggregated data.

Keywords: consistency; network meta-regression; network meta-analysis; node-splitting; inconsistency models; treatment by covariate interactions.

1. Introduction

Reviews often compare multiple treatments for the same condition. In such cases, network meta-analysis (NMA) can compare all treatments (e.g. treatment *1, 2, 3)* in a single analysis by estimating the relative treatment effects (e.g. log odds ratios) for all treatment pairings (e.g. *2* vs. *1*, *3* vs. *1*, *3* vs. *2*) using direct and indirect evidence (Higgins and Whitehead, 1996; Lu and Ades, 2004; Lu and Ades, 2006). The key assumption underlying NMA is consistency of the treatments effects across direct and indirect evidence (Lu and Ades, 2006). Many methods have been proposed to assess the consistency assumption underlying NMA (Donegan et al., 2013a), including node-splitting models (Dias et al., 2010; Van Valkenhoef et al., 2016) and inconsistency models, such as the design by treatment (DBT) inconsistency model (Higgins et al., 2012; Jackson et al., 2014; Jackson et al., 2016; Law et al., 2016; White et al., 2012) and the unrelated mean effects (URM) inconsistency model (Dias et al., 2013c).

Network meta-regression (NMR) is an extension of NMA that examines whether a covariate modifies each of the relative treatment effects (Dias et al., 2013b). A covariate may modify each relative treatment effect differently, that is, each treatment comparison may have a different covariate interaction. NMR is used to explore causes of heterogeneity or inconsistency, or when known effect modifiers exist and we wish to present results for different patient groups. Covariates may be characteristics of patients (e.g. weight), treatments (e.g. additional therapy), studies (e.g. location) or methods (e.g. allocation concealment) (Thompson and Sharp, 1999; Thompson, 1994; Thompson, 2002).

NMR results commonly consist of, for each comparison, one relative treatment effect estimated at the covariate value zero (or at the mean covariate value when the NMR model is centred) and one regression coefficient for the treatment by covariate interaction. Consistency assumptions are required for both of these parameters (Cooper et al., 2009; Donegan et al., 2013b; Donegan et al., 2012). For instance, for a three treatment NMR, where treatment *1* is taken as the reference, the consistency equation for the relative treatment effects can be written as, $d_{23} = d_{13} - d_{12}$ where for example, d_{23} is the relative treatment effect for *3* vs. *2*, and the consistency equation for the regression coefficients is $\beta_{23} = \beta_{13} - \beta_{12}$ where for example, β_{23} is the coefficient for *3* vs. *2* (Cooper et al., 2009; Dias et al., 2013b; Donegan et al., 2012). It is possible for neither assumption to hold (i.e. inconsistent relative treatment effects and inconsistent coefficients); or for only one of the assumptions to hold (i.e. either consistent relative treatment effects or consistent coefficients), which would make the results of the NMR unreliable.

Theoretically, there are eight possible scenarios that can occur when assessing whether treatment by covariate interactions exist and the consistency assumptions. Examples of the scenarios are shown in Figures 1a-1h. Each figure shows how the relative treatment effect for *3* vs. *2* changes with an increasing covariate value; separate lines are displayed for direct, indirect and all evidence. For a three treatment network, the direct evidence for *3* vs. *2* would be from trials that allocated treatments *2* and *3* and the indirect evidence for *3* vs. *2* would be from the remaining trials. Note that the lines have the same intercept when the relative treatment effects at the covariate value zero are consistent (Figure 1a-1d) and the lines have the same slope when the coefficients are consistent (Figure 1a-1b and 1e-1f). In Figure 1a, no interaction is detected using NMR and both consistency assumptions are satisfied, therefore the NMR results are valid but would not be clinically useful. On the other hand, in Figure 1b, NMR shows an interaction and both assumptions hold; therefore the NMR is reliable and could be used to draw clinical inferences. Figures 1c, 1e and 1g, show scenarios where no interaction is detected using NMR but one or more of the assumptions are not satisfied, consequently the NMR results are invalid; notably, in Figure 1c and 1g, an interaction exists when direct evidence and indirect evidence are considered separately but it is not seen when applying NMR because it is masked by the inconsistency. Lastly, in Figures 1d, 1f and 1h, an interaction is found using NMR but one or more of the assumptions do not hold so the NMR results are unreliable. The cause of inconsistency should be considered when inconsistency is found (Figures 1c-1h).

Although many methodological publications have proposed NMR analyses (Cooper et al., 2009; Dias et al., 2013b; Donegan et al., 2013b; Donegan et al., 2012; Jansen and Cope, 2012; Jansen, 2012; Nixon et al., 2007; Salanti et al., 2009; Saramago et al., 2012; Tudur Smith et al., 2007), to the authors' knowledge, no methods have been introduced for assessing the consistency assumptions underlying NMR.

In this paper, we introduce methods for assessing the consistency assumptions underlying NMR. We extend existing node-splitting models (Dias et al., 2010; Van Valkenhoef et al., 2016), the DBT inconsistency model (Higgins et al., 2012; Jackson et al., 2014; Jackson et al., 2016; Law et al., 2016; White et al., 2012) and the URM inconsistency model (Dias et al., 2013c) to incorporate treatment by covariate interactions. In section 2, we specify the NMR model and propose assessment methods that can be applied to aggregate trial-level data (i.e. trial specific relative treatment effects relative to reference arm 1 and their variances) with either continuous or categorical covariates. In section 3, we apply the methods to a real dataset and fabricated datasets illustrating key scenarios under a Bayesian framework. In section 4, we discuss the proposed methods and highlight their pros and cons.

2. Methods

We outline NMR models and then introduce methods for assessing consistency using the node-splitting models and one type of inconsistency model (i.e. URM model). New methods based on the alternative DBT inconsistency model are also presented in the supplementary material. All models are summarised in Table 1.

To set notation, let *i* denote the trial where $i = 1, \dots \dots, S$ and *S* is the number of independent trials and let *k* be the trial arm where $k = 1, \dots \dots, A_i$ and A_i is the number of arms in trial *i*. Let t_{ik} denote the treatment given in trial *i* in arm *k* where $t_{ik} \in \{1, \dots \dots, T\}$ and *T* is the number of treatments in the network. Note that treatment *1* is taken to be the reference treatment.

Suppose we have trial-level outcome data, where y_{ik} is the observed relative treatment effect (e.g. log odds ratio or mean difference) for arm k vs. arm *1* (with $k \geq 2$) in trial *i* and v_{ik} is the corresponding variance. As the relative treatment effect is a continuous measure, we assume a normal likelihood $y_{ik} \sim N(\theta_{ik,} v_{ik})$ where θ_{ik} is the mean relative treatment effect in trial *i* (with $k \geq 2$). Also, the dataset would include a study-level covariate x_i for each trial i that can be a continuous variable or an indicator variable to represent dichotomous data.

2.1. Network meta-regression models

NMR models estimate the basic regression coefficients, which are the coefficients for each treatment vs. treatment *1* (i.e. $\beta_{12}, \beta_{13}, \dots, \beta_{1T}$), and then the remaining functional coefficients (i.e. $\beta_{23}, \beta_{24}, \dots.$) are calculated as linear combinations of the basic coefficients using the consistency equations. Three NMR models have been proposed previously, each making different assumptions regarding the basic coefficients (Cooper et al., 2009; Dias et al., 2013b;

Donegan et al., 2013b; Donegan et al., 2012), that is independent (*model 1a*), exchangeable (*model 1b*) and common coefficients (*model 1c*). The decision regarding which assumption to make can be based on model fit statistics and the estimated coefficients of the models but in practice is often determined by data availability.

Model 1a can be written as

$$\theta_{ik} = \delta_{i,1k} + \beta_{t_{i1},t_{ik}} x_i$$

Where $\beta_{t_{i1},t_{ik}}$=$\beta_{1,t_{ik}}$-$\beta_{1,t_{i1}}$, $\beta_{t_{i1},t_{ik}}$ is the difference in the relative treatment effect of t_{ik} vs. t_{i1} per unit increase in the covariate x_i, or in other words, the regression coefficient for the treatment by covariate interaction. In a random-effects model, $\delta_{i,\,1k}$ (with $k \geq 2$) represents the trial-specific relative treatment effect of t_{ik} vs. t_{i1} when the covariate is zero ($x_i = 0$) and is assumed to be a realisation from a normal distribution $\delta_{i,1k} \sim N\left(d_{t_{i1},t_{ik}},\, \sigma^2\right)$ with $d_{t_{i1},t_{ik}} = d_{1,t_{ik}} - d_{1,t_{i1}}$ where $d_{t_{i1},t_{ik}}$ is the mean relative treatment effect of t_{ik} vs. t_{i1} when the covariate is zero. In a fixed-effect model, we set $\sigma^2 = 0$ to obtain $\delta_{i,\,1k}$= $d_{1,t_{ik}} - d_{1,t_{i1}}$.

M*odel 1b* is the same as *model 1a* but now $\beta_{1,t_{ik}} \sim Norm(B, \upsilon^2)$. *Model 1c* is formulated by setting $\beta_{1,t_{ik}} = \beta$ in *model 1a*; note that in this model the functional coefficients are zero because of the consistency equations (e.g. $\beta_{23} = \beta_{13} - \beta_{12} = \beta - \beta = 0$) (Cooper et al., 2009).

2.2. Assessing consistency by node-splitting

The principle aim of node-splitting models is to assess whether there is evidence of 'loop inconsistency', where loop inconsistency is defined as a difference between a result from direct and indirect evidence. Node-splitting models estimate relative treatment effects and/or regression coefficients for the interaction based on direct evidence and separate estimates from indirect evidence to explore whether they agree. Multiple node-splitting models need to be applied; one model for each comparison of interest.

To specify the node-splitting models, we extend the notation, such that the node being split is $(\hat{t}, t^*)$ where $\hat{t} \neq t^*$ and $\hat{t} < t^*$. For example, if one wants to split the node (*3*, *4*) then $\hat{t} = 3$ and $t^* = 4$.

To assess both the consistency assumptions simultaneously, node-splitting models can split the relative treatment effect and coefficient to provide, for each comparison with both direct and indirect evidence, a relative treatment effect and a coefficient estimated from direct evidence and an effect and coefficient based on indirect evidence. The model that splits the relative treatment effect and coefficient and includes independent interactions (*model 2.1a*) is an extension of *model 1a* as follows:

$$\theta_{ik} = \begin{cases} \delta_{i,1k} + \beta_{t_{i1},t_{ik}} x_i & if\ \ t_{i1} \neq \hat{t} \text{ and/or } t_{ik} \neq t^* \\ \delta_{i,1k} + \beta^{dir} x_i & if\ \ t_{i1} = \hat{t} \text{ and } t_{ik} = t^* \end{cases}$$

Where $\beta_{t_{i1},t_{ik}}$=$\beta_{1,t_{ik}}$-$\beta_{1,t_{i1}}$, $\beta_{t_{i1},t_{ik}}$ represents the difference in the relative treatment effect of t_{ik} vs. t_{i1} per unit increase in the covariate estimated using indirect evidence, and β^{dir} represents the difference in the relative treatment effect of t^* vs. $\hat{t}$ per unit increase in the covariate estimated using direct evidence. In a random-effects model, if trial *i* allocated $t^*$and $\hat{t}$, that is, $t_{i1} = \hat{t}$ and $t_{ik} = t^*$, then $\delta_{i,1k} \sim N\left(d^{dir}, \sigma^2\right)$ where d^{dir} represents the mean relative treatment effect of t^*vs. $\hat{t}$ when the covariate value is zero estimated using direct evidence; whereas if trial *i* did not allocate $t^*$and $\hat{t}$, that is, $t_{i1} \neq \hat{t}$ and/or $t_{ik} \neq t^*$, then $\delta_{i,1k} \sim N\left(d_{t_{i1},t_{ik}}, \sigma^2\right)$ where $d_{t_{i1},t_{ik}}$ represents the mean relative treatment effect of t_{ik} vs. t_{i1} when the covariate value is zero estimated using indirect evidence and $d_{t_{i1},t_{ik}} = d_{1,t_{ik}} - d_{1,t_{i1}}$.

To assess only the consistency of the relative treatment effects, node-splitting models can split the relative treatment effect alone to produce a single coefficient that is estimated using all evidence and two relative treatment effects (i.e. one estimated using direct evidence and the other estimated using the indirect evidence). The model that splits the relative treatment effect alone and includes independent interactions (*model 2.2a*) is

$$\theta_{ik} = \delta_{i,1k} + \beta_{t_{i1},t_{ik}} x_i$$

where $\beta_{t_{i1},t_{ik}}$ represents the difference in the relative treatment effect of t_{ik} vs. t_{i1} per unit increase in the covariate estimated using all evidence. In this model, the trial-specific relative treatment effects, $\delta_{i,\,1k}$ are distributed in the same way as in *model 2.1a*.

Likewise, to assess the consistency of the coefficients alone, a node-splitting model can split only the coefficient to estimate a single relative treatment effect using all evidence and two coefficients (i.e. one estimated from direct evidence and the other from indirect evidence). The model that splits only the coefficient and includes independent interactions (*model 2.3a*) is the same as *model 2.1a* except the trial-specific relative treatment effects, $\delta_{i,\,1k}$ are distributed as $\delta_{i,1k} \sim N\left(d_{t_{i1},t_{ik}},\, \sigma^2\right)$ where $d_{t_{i1},t_{ik}}$ represents the mean relative treatment effect of t_{ik} vs. t_{i1} when the covariate value is zero estimated using all evidence.

Node-splitting models can be adapted to include exchangeable (*models 2.1b, 2.2b, 2.3b*) or common (*models 2.1c 2.2c, 2.3c*) interactions as described in section 2.1. Note that *model 2.1c* and *2.3c* fix each functional coefficient based on indirect evidence (i.e. $\beta_{t_{i1},t_{ik}}$when $t_{i1} \neq 1$) to be zero whereas the corresponding result from direct evidence (β^{dir}) is not.

The level of consistency can be assessed, by comparing the model fit of the NMR (*model 1(a, b, or c)*) with that of the node-splitting models (*models 2.1(a, b, or c), 2.2(a, b, or c),* and *2.3(a, b, or c)*); inconsistency is indicated if a node-splitting model is an improved fit. Moreover, if the between trial variance is lower in the node-splitting models as compared to the NMR, inconsistency may exist. Also, for each treatment comparison, the size, direction, and precision of the relative treatment effect estimated using direct evidence can be compared with that estimated using indirect evidence. Such comparisons are subjective and when results are presented graphically and compared, care must be taken because the scale and shape of the plots can affect how different the results appear to be. Furthermore, when using Bayesian methods, for each comparison, the probability that the direct and indirect evidence differs can be calculated. For each treatment pairing, the inconsistency estimate (*IE*), that is the difference between the relative treatment effect from direct evidence and indirect evidence can be calculated at each iteration of the chain, and the number of iterations for which $IE \geq 0$ is counted. It is then possible to calculate the probability (*prob*) that the relative treatment effect from direct evidence exceeds the relative treatment effect from indirect evidence, by dividing the number of counted iterations by the total number of iterations of the chain. Lastly, assuming that the posterior distribution of the difference (IE) is symmetric and unimodal, the probability that the direct and indirect evidence agree is given by $P = 2 \times minimum(prob, 1 - prob)$ (Dias et al., 2010; Marshall and Spiegelhalter, 2007). Likewise, the regression coefficients from direct and indirect evidence can be compared in the same way.

2.3. Assessing consistency using URM models.

URM models assess global consistency, which is inconsistency somewhere in the treatment network, by comparing the results from an NMR model with those from an URM model (Dias et al., 2013c).

The URM model that assesses the consistency of the relative treatment effects and coefficients and includes independent interactions (*model 3.1a*) is the same as the NMR model (*model 1a*) but it does not incorporate the consistency equations (i.e. $d_{t_{i1},t_{ik}} = d_{1,t_{ik}} - d_{1,t_{i1}}$ and $\beta_{t_{i1},t_{ik}}$=$\beta_{1,t_{ik}}$-$\beta_{1,t_{i1}}$), and as such, the model parameters are estimated using direct evidence only. *Model 3.1a* is equivalent to fitting separate pair-wise meta-regressions, except, *model 3.1a* assumes the between trial variance (σ^2) is equal across comparisons but the pair-wise meta-regressions would not.

The URM model that assesses only consistency of the relative treatment effects and includes independent interactions (*model 3.2a*) is the same as model 3*.1a* but incorporates the consistency equation for the coefficients. Likewise, the UMR model that assesses only consistency of the coefficients with independent interactions (*model 3.3a*) is same as model 3*.1a* but includes the consistency equation for the relative treatment effects.

Exchangeable (*models 3.1b, 3.2b, 3.3b*) or common (*models 3.1c, 3.2c, 3.3c*) interactions can be included. However, it is worth noting that the independent, exchangeable or common assumptions are slightly different to those specified for the NMR models (*models 1a, 1b* and *1c*). In the NMR models, we assume the basic regression coefficients (i.e. $\beta_{12}, \beta_{13}, \dots, \beta_{1T}$) are independent, exchangeable or common. However, when the consistency equation for the coefficients is not used in the URM model (i.e. *models 3.1(a, b, or c)* and *3.3(a, b, or c)*), we can assume that all regression coefficients, that is basic and functional coefficients, are independent, exchangeable (i.e. $\beta_{t_{i1},t_{ik}} \sim Norm(B, \upsilon^2)$) or common (i.e. $\beta_{t_{i1},t_{ik}} = \beta$). In particular, this means that when including common interactions, the functional coefficients in the NMR model (*model 1c*) are forced to be zero but this is not so in the URM model (*models 3.1c* and *3.3c*).

To determine consistency, the model fit of the NMR model (*model 1(a, b, or c)*) and the fit of the URM models (*models 3.1(a, b, or c), 3.2(a, b, or c)* and *3.3(a, b, or c)*) can be compared; when an URM model is an improved fit, inconsistency may be present. Also, differences between the relative treatment effects and regression coefficients produced from the NMR model and those from the URM models may suggest inconsistency.

2.4. Including multi-arm trials

The models can be applied to datasets including multi-arm trials providing that the correlation between the observed relative treatment effect (y_{ik}) and the trial-specific relative treatment effects ($\delta_{i,1k}$) is taken into account. For each multi-arm trial *i* with *m* arms, the observed relative treatment effects and the trial-specific relative treatment effects are assumed to follow multivariate normal distributions

$$\begin{pmatrix} y_{i2} \\ \vdots \\ y_{im} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_{i2} \\ \vdots \\ \theta_{im} \end{pmatrix}, \begin{pmatrix} v_{i2} & \dots & cov(y_{i2}, y_{im}) \\ \vdots & \ddots & \vdots \\ cov(y_{i2}, y_{im}) & \dots & v_{im} \end{pmatrix}\right)$$

and

$$\begin{pmatrix} \delta_{i,\,12} \\ \vdots \\ \delta_{i,\,1m} \end{pmatrix} \sim N\left(\begin{pmatrix} d_{1,t_{i2}} - d_{1,t_{i1}} \\ \vdots \\ d_{1,t_{im}} - d_{1,t_{i1}} \end{pmatrix}, \begin{pmatrix} \tau^2 & \dots & \tau^2/2 \\ \vdots & \ddots & \vdots \\ \tau^2/2 & \dots & \tau^2 \end{pmatrix}\right).$$

Furthermore, there is an extra consideration when fitting node-splitting models (Dias et al., 2010; Van Valkenhoef et al., 2016). If one wants to split node (t_{i1}, t_{ik}) then a multi-arm trial will contribute direct evidence to the relative treatment effect (d^{dir}) as required because $\hat{t} = t_{i1}$. However, the multi-arm trial would not contribute direct evidence to the estimation of the relative treatment effect, d^{dir}, if one splits another node (e.g. t_{i2}, t_{i3}) because $\hat{t} \neq t_{i1}$. Therefore, to overcome this problem, when a multi-arm trial compared the two treatments t^* and $\hat{t}$, in addition to other treatments, treatment $\hat{t}$ is taken to be the baseline treatment t_{i1} for that study.

Note that for URM models including multi-arm trial data, the URM model is not the same as fitting separate pair-wise meta-regressions because the correlation in multi-arm trials is taken into account but would not be in pair-wise analyses; also, the URM model only uses t_{i1} as the baseline treatment so direct evidence for some pairwise comparisons would not be used whereas pairwise meta-regression could utilise all direct evidence.

3. Application to datasets

3.1. Datasets

Here, the methods proposed in section 2 are applied to a real dataset and four fabricated datasets that have been manipulated to demonstrate specific scenarios.

3.1.1. Malaria dataset

Two Cochrane reviews and the corresponding trials were used to construct the malaria dataset; reviews compared artemether (AR), quinine (QU) and artesunate (AS) (Esu et al., 2014; Sinclair et al., 2012). Randomised controlled trials including patients with severe malaria were eligible. Age was considered to be an effect modifier because the clinical features of malaria differ by age and thus all treatment recommendations are stratified by age in the reviews and WHO treatment guidelines (World Health Organisation, 2015). Event rates for the primary outcome, death, and the covariate, average age of patients in each trial, was extracted. Two studies with missing covariate data were deleted from the dataset. Using the event rates, trial-specific log odds ratios and their standard deviations were calculated in R. Table S1 displays the data. Figure 2 shows the network diagram.

3.1.2. Fabricated datasets

Four fabricated datasets were constructed by manipulating the malaria dataset to illustrate key scenarios: (1) no interaction present and the relative treatment effects and regression coefficients are consistent (Figure 1a); (2) interaction exists and the relative treatment effects and coefficients are consistent (Figure 1b) (3) interaction exists and the relative treatment effects are consistent but the coefficients are inconsistent (Figure 1d); (4) no interaction present and the relative treatment effects are consistent but the coefficients are inconsistent (Figure 1g). Example R code to generate the datasets is given in the supplementary material.

Analogous to the malaria dataset, each dataset compared three treatments (AS, AR, QU), there was direct evidence for each possible comparison, no multi-arm trials contributed, and a dichotomous outcome and continuous covariate was of interest. Ten trials contributed direct evidence to each comparison. For each study, a continuous covariate was taken to be a realisation from Normal distribution (i.e. $N(17, 10^2)$) truncated at zero to ensure the covariate values were similar to those observed in the malaria dataset.

The log odds ratios and regression coefficients were chosen to be similar to those estimated in the original dataset. For each dataset, the log odd ratio at zero covariate of trials comparing treatments AR and AS was 0.2, trials comparing treatments QU and AS was 0.23, and trials of treatments QU and AR was 0.03. For dataset one, the coefficient for each comparison was zero. For dataset two, the coefficient for trials comparing treatments AR and AS was 0.02, trials comparing treatments QU and AS was 0.02, and trials of treatments QU and AR was 0. For dataset three, the coefficient for trials comparing treatments AR and AS was 0.01, trials of treatments QU and AS was 0.04, and trials comparing treatments QU and AR was 0. For dataset four, the coefficient for trials comparing treatments AR and AS was -0.04, trials of treatments QU and AS was 0.04, and trials of treatments QU and AR was 0.

The trial-specific observed log odds ratios were estimated from the values of log odds ratio at zero covariate, the coefficients, and the covariates. The between trial variance was zero. The standard error of the observed log odds ratio was 0.2 for each trial.

3.2. Implementation

All models were fitted to the datasets using WinBUGS 1.4.3 and the R2WinBUGS package in R. Example code is provided as supplementary material. For the malaria dataset, all models in Table 1 were fitted. For the fabricated datasets, only fixed-effect versions of *models 1a*, *2.1a*, *3.1a* and *4.1a* were applied because the between trial variance was zero and the coefficients differed across comparisons. See Table S2 for the parameterisation of the DBT models. The covariates were centred at their mean. All parameters were given non-informative normal prior distributions (i.e. $N(0, 100000)$) except the between-trial standard deviation that was assumed to follow a non-informative uniform distribution (i.e. $Uni(0, 10)$) and a weakly informative prior distribution (i.e. $uniform\,(0, 2)$) was specified for the standard deviation of the exchangeable regression coefficients. Three chains with different initial values were run for 300,000 iterations. The initial 100,000 draws were discarded and chains were thinned such that every fifth iteration was retained. Convergence of the chains was assessed by inspecting trace plots of the draws.

Model fit and complexity of models was assessed using the deviance information criterion (DIC) defined as $DIC = \overline{D} + p_D$ where $\overline{D}$ is the posterior mean of the residual deviance and p_D is the effective number of parameters (Spiegelhalter et al., 2002). A model with a smaller DIC was preferable to a model with a larger DIC but differences of less than three units were not considered meaningful. When models had little difference in DIC, the simplest model was chosen.

3.3. Results

Results from NMR, node-splitting and URM models are presented here. The results from DBT models are presented in supplementary material.

3.3.1. Malaria dataset

NMR models

Comparing fixed-effect and random-effect NMR models (*models 1a*, *1b*, *1c*), the DICs from all NMR models variations are similar (DICs 24.93-26.76 in Table S3). Also, the estimated regression coefficients for the treatment by average age interactions were quite similar for each model variation (Table S4). Therefore, results from the simplest model, the fixed-effect NMR with common interactions (*model 1c*) are presented.

The results of *model 1c* show that there is evidence of a small interaction between relative treatment effect and average age for AR vs. AS and QU vs. AS; the posterior median of the common regression coefficient for AR vs. AS and QU vs. AS is 0.0132 with 95% credibility interval (CrI) (0.0018, 0.0244) (Table S4). There is no interaction for QU vs. AR because the model fixes the coefficient to be zero. However, before using these results to draw clinical inferences, the underlying consistency assumptions must be assessed.

Node-splitting models

Table 2 shows model fit assessment results for fixed-effect node-splitting models with common interactions (*models 2.1c, 2.2c, 2.3c*). The DIC of the NMR model (DIC=25.29) is similar to those of the node-splitting models (DICs 23.75-27.95) indicating that the model is not improved by splitting each node, lending support to the consistency assumptions.

The results from node-splitting are displayed in Table 3. In the model that assesses consistency of both the log odds ratio and the coefficient (*model 2.1c*), the log odds ratios for AR vs. AS (-2.3540 95% CrI (-6.7650, 2.0530)) and QU vs. AS (0.4316 95% CrI (0.2833, 0.5797)) based on direct evidence differs with those from indirect evidence (i.e. 0.1985 95% CrI (-0.0815, 0.4782) and -2.1000 95% CrI (-6.4180, 2.4430) respectively) because only two trials contribute direct evidence for AR vs. AS and therefore the results are influenced by the vague prior distribution. A similar, but less pronounced, inconsistency is also seen for the corresponding coefficients. Yet, the probability of agreement between direct and indirect evidence is low for the coefficient for QU vs. AR (P=0.06) but not remarkably low for other comparisons or the log odds ratios (Ps 0.24-0.77). Similar conclusions are drawn from models that split either the log odds ratio or the regression coefficient only (*models 2.2c* and *2.3c*). The consistency of the direct and indirect evidence is also supported graphically in Figure 3, which displays the posterior distributions of the centred log odds ratios and regression coefficients and in Figure 4, where the log odds ratio versus average age is plotted.

URM models

Table 2 also displays model fit assessment results for fixed-effect URM models with common interactions (*models 3.1c*, *3.2c*, *3.3c*). The DIC of the NMR model (DIC=25.29) is similar to those from the URM models the assess consistency of both the log odds ratio and coefficient (DIC=23.94) or the log odds ratio alone (DIC= 27.27) (*models 3.1c and 3.2c*) but is slightly higher than that from the model that assesses the coefficient alone (DIC=21.96) (*model 3.3c*) indicating a possible inconsistency on a coefficient.

See Table 4 for the results from the NMR model and URM models. The results from the URM models are quite similar to those from the NMR model with the exception of the regression coefficient for QU vs. AR. This difference in the coefficient for QU vs. AR is because of the different assumptions underlying the two models; the NMR model sets the regression coefficients for AR vs. AS and QU vs. AS to be identical (i.e. 0.0132 95% CrI (0.0018, 0.0244)) and the coefficient for QU vs. AR to be zero, whereas all three coefficients are set to be identical in the URM model (i.e. 0.0145 95% CrI (0.0044, 0.0247)).

Overall, there is evidence of an interaction from the NMR but also evidence of inconsistency; the node-splitting models show evidence of loop inconsistency for the coefficient of QU vs. AR and the URM models support this showing a possible inconsistency of the coefficients.

3.3.2. Fabricated datasets

Dataset 1: no interaction and consistency.

The DICs from each model (*models 1a*, *2.1a*, *3.1a*) are similar (8.01-12.00) therefore there is no obvious sign of inconsistency (Table 5). Using the results from node-splitting (*model 2.1a*), the log odds ratios and coefficients based on direct and indirect evidence are very similar and the probabilities of agreement between direct and indirect evidence are practically one (Table 6). The results from the NMR model are also similar to those from the URM model (*model 3.1a*) (Table 7) indicating consistency. Overall, the NMR model does not show that a treatment by average age interaction exists (Table 7) and there is no evidence of loop inconsistency using node-splitting, or global inconsistency using the URM model. Figure 5, which shows the results from the NMR model and node-splitting models, supports this conclusion.

Dataset 2: interaction and consistency.

The DICs from the models (*models 1a, 2.1a, 3.1a*) are again similar (8.00-11.99) indicating consistent evidence (Table 5). From node-splitting (*model 2.1a*), the log odds ratios and the coefficients based on direct and indirect evidence are almost identical and the probabilities of agreement of direct and indirect evidence are practically one (Table 6); Figure 5 shows the results graphically. The URM model (*model 3.1a*) also gives comparable results to the NMR model (Table 7). In conclusion, the NMR model shows that an interaction exists for AR vs. AS (0.0200 95% CrI (0.0074, 0.0327)) and QU vs. AS (0.0200 95% CrI (0.0080, 0.0321)) (Table 7) and there is no loop inconsistency using node-splitting, or global inconsistency using the URM model.

Dataset 3: interaction and inconsistency.

The DIC from the NMR model (*model 1a*) (DIC=47.14) is much higher than those from node-splitting (*model 2.1a*) and the URM model (*model 3.1a*) (11.97-11.99) suggesting inconsistency (Table 5). From node-splitting, the log odds ratios based on direct and indirect evidence are comparable but the coefficients for AR vs. AS (0.0100 95% CrI (-0.0039, 0.0241)) and QU vs. AS (0.0400 95% CrI (0.0298, 0.0503)) and QU vs. AR (0.0000 95% CrI (-0.0125, 0.0126)) from direct evidence differ from those from indirect evidence (i.e. 0.0400 95% CrI (0.0237, 0.0562), 0.0099 95% CrI (-0.0088, 0.0289), and 0.0300 95% CrI (0.0127, 0.0474) respectively); the probabilities of agreement of direct and indirect evidence are very high (Ps 0.9982-0.9990) for the log odds ratios and very low for the coefficients (Ps 0.0057-0.0062) (Table 6). The URM model also gives results that differ somewhat from those of the NMR model (see Table 7). To summarise, the NMR model shows that an interaction exists for AR vs. AS (0.0187 95% CrI (0.0082, 0.0292)), QU vs. AS (0.0335 95% CrI (0.0244, 0.0425)) and QU vs. AR (0.0147 95% CrI (0.0047, 0.0248)) (Table 7) but there is also loop inconsistency in the size of the underlying coefficients based on direct and indirect evidence that is seen using node-splitting (Figure 5); the URM model identifies global inconsistency.

Dataset 4: no interaction and inconsistency.

The DIC from the NMR model (*model 1a*) (DIC=188.36) is much higher than those from node-splitting (*model 2.1a*) and the URM model (*model 3.1a*) (11.99-12.00) indicating inconsistency (Table 5). Similar to dataset 3, in node-splitting models, the log odds ratios based on direct and indirect evidence are comparable but the coefficients for AR vs. AS (-0.0400 95% CrI (-0.0553, -0.0246)) and QU vs. AS (0.0400 95% CrI (0.0273, 0.0529)) and QU vs. AR (0.0000 95% CrI (-0.0115, 0.0116)) from direct evidence differ from those from indirect evidence (i.e. 0.0399 95% CrI (0.0227, 0.0574), -0.0400 95% CrI (-0.0591, -0.0208), and 0.0800 95% CrI (0.0600, 0.1000) respectively); the probabilities of agreement of direct and indirect evidence are very high for log odds ratios (Ps 0.9976-1.000) and zero for the coefficients (Table 6). Also, results from the URM model are different from those of the NMR model (see Table 7). Overall, the NMR model shows that no interaction exists (Table 7) but there is inconsistency in the direction of the underlying coefficients based on direct and indirect evidence and this trend can be seen using node-splitting (Figure 5); the URM model suggests global inconsistency respectively but these models cannot show the underlying trend.

4. Discussion

We have shown that node-splitting and inconsistency models can be useful for assessing the underlying consistency assumptions of NMR when using aggregate data. Once consistency has been assessed, the analyst must decide which results to present. If the direct and indirect evidence are consistent, the results from the NMR should be reliable. However, the level of heterogeneity (from the NMR or standard pair-wise analyses) and goodness of fit of the NMR should be considered when drawing conclusions from the results. If there is inconsistency, the results from the NMR are questionable and the causes of inconsistency should be considered. In some scenarios, for example, when inconsistency masks an interaction, as shown in Figure 1c and 1g, the results would not be useable. If the original purpose of the NMR was to explore causes of heterogeneity or inconsistency in an NMA and there is no interaction and no inconsistency masking interactions in the NMR, then analysts could proceed by exploring other potentially relative treatment effect modifying covariates or reconsidering the eligibility criteria.

Each of the proposed methods has different pros and cons. DBT models assess design and loop consistency and can assess global inconsistency, while node-splitting assesses loop consistency and URM models assess global inconsistency; loop inconsistency is well recognised in the methodological literature but design consistency is a newer concept (Higgins et al., 2012; White et al., 2012). Furthermore, the DBT model requires parameterisation by the analyst therefore, the analyst needs to have a good understanding of the model and parameters. Key advantages of the DBT model and node-splitting is that inconsistency estimates and the probability that direct and indirect evidence agree can be obtained; however, the URM model does not provide such results. Moreover, concerns regarding multiple testing may apply to node-splitting and the DBT models where probabilities are calculated, particularly when a Frequentist approach is taken; therefore, it is important to compare model fit statistics across models, and also to be cautious in interpreting 'p-values' making sure to allow for multiple testing. One disadvantage of node-splitting is that, as one model is fitted for every comparison with contributing direct and indirect evidence, many models may need to be fitted which is computationally demanding; whereas only one inconsistency model would need to be applied.

Ideally, all three approaches (i.e. node-splitting, DBT model, URM model) would be applied to provide a thorough assessment of consistency. However, in practice, the reviewer may select their preferred approach depending on the ease of application in software etc. We recommend that at least one of the global tests (i.e. inconsistency models) and also node-splitting are performed. Our preference is node-splitting because estimates from direct and indirect evidence can be found.

We proposed and applied methods to trial-level aggregated data in this article. However, it is straightforward to adapt the models to accommodate any type of arm-level outcome data, that is, a summary of the outcome data for each arm of each trial and a covariate value for each trial. To adapt the models, a suitable link function would be chosen and nuisance parameters are included in the model to represent the effect of the baseline treatment in arm *1* of trial *i*. Further details regarding arm-level network meta-analysis models are given by Dias et al (Dias et al., 2013a)

Moreover, collection and use of individual patient data is generally advantageous over aggregate data when studying patient-level covariates because they avoid ecological biases (Riley et al., 2008; Riley and Steyerberg, 2010). Yet, it is more common to explore patient-level covariates (e.g. patient age) using study-level covariate summaries (e.g. average age of patients) in meta-regression such as in the malaria dataset. However, when using aggregate data, the possibility of confounding and ecological biases should be considered when patient-level covariates are explored.

There are a number of issues that can arise when applying the methods, particularly with aggregate data. Parameter estimation can be a problem with limited data, such that models cannot be fitted at all, interactions exist but cannot be detected, or inconsistency exists but is not found. For instance, when all the trials that contribute to the estimation of a regression coefficient have the same covariate value or when only one trial contributes to a coefficient, this would preclude the use of models with independent interactions but analysts may be able to apply an model with exchangeable or common interactions providing studies that contribute to another basic coefficient have different covariate values. For example, when exploring an interaction between relative treatment effect and study location (i.e. continent), studies that contribute to results for comparison *2* vs. *1* may all be carried out on the same continent provided that studies that contribute to comparison *3* vs. *1* are located on different continents. Parameter estimation may particularly be a problem when fitting the DBT model because the inconsistency estimates would be imprecise when the number of trials in one or more designs is limited; to overcome this one could assume exchangeability of the inconsistency factors or use informative prior distributions. Similarly, if direct evidence is limited for some comparisons (i.e. few trials or covariate values), the URM model and node-splitting models would produce imprecise results and informative prior distributions may need to be used. Ideally any informative prior distributions would be evidence-based by eliciting them from similar meta-analyses or experts' beliefs. Finally, it is also worth emphasising that no evidence of inconsistency does not automatically imply there is consistency; inconsistency may exist but cannot be detected when data are limited and results are imprecise and therefore arguably the consistency assumptions and the NMR results are questionable. In the same way, in such cases, no evidence of a treatment by covariate interaction does not imply there is truly no interaction.

Conversely, with abundant data, additional modelling extensions may be feasible. For example, in node-splitting models, we have assumed the between trial variance is the same for direct evidence and indirect evidence, yet it is possible to incorporate two variances, one of each type of evidence. Also, the models could be adapted to include more than one covariate or other variance structures (Lu and Ades, 2009).

In conclusion, consistency of the assumptions underlying NMR must be assessed when NMR is applied, even when no treatment by covariate interactions are detected. It is possible that inconsistency is masking an interaction. Furthermore, results of an NMR should not be reported without assessing the underlying assumptions to determine whether the results are valid and reliable.

Acknowledgements

This research was funded by the Medical Research Council (http://www.mrc.ac.uk/, grant number MR/K021435/1) as part of a career development award in biostatistics awarded to SDo. We are grateful to the two anonymous peer reviewers for their helpful comments.

References

COOPER, N., SUTTON, A., MORRIS, D., ADES, A. & WELTON, N. 2009. Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: Application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. *Stat Med,* 28**,** 1861-1881.

DIAS, S., SUTTON, A. J., ADES, A. E. & WELTON, N. J. 2013a. Evidence Synthesis for Decision Making 2: A Generalized Linear Modeling Framework for Pairwise and

Network Meta-analysis of Randomized Controlled Trials. *Med Decis Making,* 33**,** 607-617.

DIAS, S., SUTTON, A. J., WELTON, N. J. & ADES, A. E. 2013b. Evidence Synthesis for Decision Making 3: Heterogeneity—Subgroups, Meta-Regression, Bias, and Bias-Adjustment. *Med Decis Making,* 33**,** 618-640.

DIAS, S., WELTON, N. J., CALDWELL, D. M. & ADES, A. E. 2010. Checking consistency in mixed treatment comparison meta-analysis. *Stat Med,* 29**,** 932-944.

DIAS, S., WELTON, N. J., SUTTON, A. J., CALDWELL, D. M., LU, G. & ADES, A. E. 2013c. Evidence Synthesis for Decision Making 4: Inconsistency in Networks of Evidence Based on Randomized Controlled Trials. *Med Decis Making,* 33**,** 641-656.

DONEGAN, S., WILLIAMSON, P., D'ALESSANDRO, U. & TUDUR SMITH, C. 2013a. Assessing key assumptions of network meta-analysis: a review of methods. *Res Syn Meth*, 4**,** 291-323.

DONEGAN, S., WILLIAMSON, P., D'ALESSANDRO, U., GARNER, P. & TUDUR SMITH, C. 2013b. Combining individual patient data and aggregate data in mixed treatment comparison meta-analysis: Individual patient data may be beneficial if only for a subset of trials. *Stat Med,* 32**,** 914-930.

DONEGAN, S., WILLIAMSON, P., D'ALESSANDRO, U. & TUDUR SMITH, C. 2012. Assessing the consistency assumption by exploring treatment by covariate interactions in mixed treatment comparison meta-analysis: individual patient-level covariates versus aggregate trial-level covariates. *Stat Med*, 31**,** 3840-3857.

ESU, E., EFFA, E. E., OPIE, O. N., UWAOMA, A. & MEREMIKWU, M. M. 2014. Artemether for severe malaria. *Cochrane Database Syst Rev,* 9, CD010678..

HIGGINS, J. & WHITEHEAD, A. 1996. Borrowing strength from external trials in a meta-analysis. *Stat Med,* 15**,** 2733-49.

HIGGINS, J. P. T., JACKSON, D., BARRETT, J. K., LU, G., ADES, A. E. & WHITE, I. R. 2012. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Res Syn Meth,* 3**,** 98–110.

JACKSON, D., BARRETT, J. K., RICE, S., WHITE, I. R. & HIGGINS, J. P. T. 2014. A design-by-treatment interaction model for network meta-analysis with random inconsistency effects. *Stat Med,* 33**,** 3639-3654.

JACKSON, D., BODDINGTON, P. & WHITE, I. R. 2016. The design-by-treatment interaction model: a unifying framework for modelling loop inconsistency in network meta-analysis. *Res Syn Meth,* 7**,** 329-332.

JANSEN, J. & COPE, S. 2012. Meta-regression models to address heterogeneity and inconsistency in network meta-analysis of survival outcomes. *BMC Med Res Methodol,* 12**,** 152.

JANSEN, J. P. 2012. Network meta-analysis of individual and aggregate level data. *Res Syn Meth,* 3**,** 177-190.

LAW, M., JACKSON, D., TURNER, R., RHODES, K. & VIECHTBAUER, W. 2016. Two new methods to fit models for network meta-analysis with random inconsistency effects. *BMC Med Res Methodol,* 16**,** 87.

LU, G. & ADES, A. 2004. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med,* 23**,** 3105 - 3124.

LU, G. & ADES, A. 2006. Assessing evidence inconsistency in mixed treatment comparisons. *J Am Stat Assoc* 101**,** 447-459.

LU, G. & ADES, A. 2009. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics,* 10**,** 792-805.

MARSHALL, E. & SPIEGELHALTER, D. 2007. Identifying outliers in Bayesian hierarchical models: a simulation-based approach. *Bayesian Anal,* 2**,** 409-444.

NIXON, R. M., BANSBACK, N. & BRENNAN, A. 2007. Using mixed treatment comparisons and meta-regression to perform indirect comparisons to estimate the efficacy of biologic treatments in rheumatoid arthritis. *Stat Med,* 26**,** 1237-54.

RILEY, R. D., LAMBERT, P. C., STAESSEN, J. A., WANG, J., GUEYFFIER, F., THIJS, L. & BOUTITIE, F. 2008. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Stat Med,* 27**,** 1870-1893.

RILEY, R. D. & STEYERBERG, E. W. 2010. Meta-analysis of a binary outcome using individual participant data and aggregate data. *Res Syn Meth,* 1**,** 2-19.

SALANTI, G., MARINHO, V. & HIGGINS, J. P. T. 2009. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. *J Clin Epidemiol,* 62**,** 857-864.

SARAMAGO, P., SUTTON, A. J., COOPER, N. J. & MANCA, A. 2012. Mixed treatment comparisons using aggregate and individual participant level data. *Stat Med,* 31**,** 3516-3536.

SINCLAIR, D., DONEGAN, S., ISBA, R. & LALLOO DAVID, G. 2012. Artesunate versus quinine for treating severe malaria. *Cochrane Database Syst Rev,* 6, *CD005967*.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. & VAN DER LINDE, A. 2002. Bayesian measures of model complexity and fit. *Med Decis Making,* 64**,** 583-639.

THOMPSON, S. & SHARP, S. 1999. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med,* 18**,** 2693 - 2708.

THOMPSON, S. G. 1994. Systematic Review: Why sources of heterogeneity in meta-analysis should be investigated. *BMJ,* 309**,** 1351-1355.

THOMPSON, S. G., HIGGINS, J. P. T. 2002. How should meta-regression analyses be undertaken and interpreted? *Stat Med,* 21**,** 1559-1573.

TUDUR SMITH, C., MARSON, A., CHADWICK, D. & WILLIAMSON, P. 2007. Multiple treatment comparisons in epilepsy monotherapy trials. *Trials,* 8**,** 34.

VAN VALKENHOEF, G., DIAS, S., ADES, A. E. & WELTON, N. J. 2016. Automated generation of node-splitting models for assessment of inconsistency in network meta-analysis. *Res Syn Meth,* 7**,** 80-93.

WHITE, I. R., BARRETT, J. K., JACKSON, D. & HIGGINS, J. P. T. 2012. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. *Res Syn Meth,* 3**,** 111-125.

WORLD HEALTH ORGANISATION 2015. Guidelines for the treatment of malaria. Third edition ed.

		Models including independent treatment by covariate interactions	**Models including exchangeable treatment by covariate interactions**	**Models including common treatment by covariate interactions**
NMR models		Model 1a	Model 1b	Model 1c
Node-splitting models	**Models splitting the relative treatment effect and the regression coefficient for the interaction.**	Model 2.1a	Model 2.1b	Model 2.1c
	Models splitting the relative treatment effect only.	Model 2.2a	Model 2.2b	Model 2.2c
	Models splitting the regression coefficient for the interaction only.	Model 2.3a	Model 2.3b	Model 2.3c
URM models	**Models assessing consistency of the relative treatment effect and the regression coefficient for the interaction.**	Model 3.1a	Model 3.1b	Model 3.1c
	Models assessing consistency of the relative treatment effect only.	Model 3.2a	Model 3.2b	Model 3.2c
	Models assessing consistency of the regression coefficient for the interaction only.	Model 3.3a	Model 3.3b	Model 3.3c
DBT models	**Models assessing consistency of the relative treatment effect and the regression coefficient for the interaction.**	Model 4.1a	Model 4.1b	Model 4.1c
	Models assessing consistency of the relative treatment effect only.	Model 4.2a	Model 4.2b	Model 4.2c
	Models assessing consistency of the regression coefficient for the interaction only.	Model 4.3a	Model 4.3b	Model 4.3c

Table 1: Proposed model variations.

DBT: design by treatment; NMR: network meta-regression; URM: unrelated mean effects.

Model	**Mean residual deviance**	**p_D**	**DIC**
NMR model (*model 1c*)	22.29	3.00	25.29
Node-splitting model splitting the log odds ratio and regression coefficient: AR vs. AS (*model 2.1c*)	22.97	4.99	27.95
Node-splitting model splitting the log odds ratio and regression coefficient: QU vs. AS (*model 2.1c*)	22.96	4.98	27.93
Node-splitting model splitting the log odds ratio and regression coefficient: QU vs. AR (*model 2.1c*)	20.65	5.00	25.65
Node-splitting model splitting the log odds ratio only: AR vs. AS (*model 2.2c*)	23.27	4.01	27.27
Node-splitting model splitting the log odds ratio only: QU vs. AS (*model 2.2c*)	23.27	4.01	27.29
Node-splitting model splitting the log odds ratio only: QU vs. AR (*model 2.2c*)	23.27	4.01	27.27
Node-splitting model splitting the regression coefficient only: AR vs. AS (*model 2.3c*)	23.19	4.01	27.19
Node-splitting model splitting the regression coefficient only: QU vs. AS (*model 2.3c*)	23.19	4.01	27.19
Node-splitting model splitting the regression coefficient only: QU vs. AR (*model 2.3c*)	19.74	4.01	23.75
URM model assessing consistency of the log odds ratio and regression coefficient (*model 3.1c*)	19.93	4.01	23.94
URM model assessing consistency of the log odds ratio only (*model 3.2c*)	23.27	4.01	27.27
URM model assessing consistency of the regression coefficient only (*model 3.3c*)	18.96	3.00	21.96

Table 2: Model fit assessment results for fixed-effect models with common treatment by average age interactions for the malaria dataset.
Number of data points: 24
AR: artemether; AS: artesunate; DIC: deviance information criterion; QU: quinine; NMR: network meta-regression; URM: unrelated mean effects.

Model type	Parameter	Evidence	Posterior median (95% credibility interval), P		
			AR vs. AS	**QU vs. AS**	**QU vs. AR**
Splitting the log odds ratio and regression coefficient (*model 2.1c*)	**Log odds ratio (centred)**	**Direct**	-2.3540 (-6.7650, 2.0530)*	0.4316 (0.2833, 0.5797)	0.2882 (0.0449, 0.5315)
		Indirect	0.1985 (-0.0815, 0.4782)	-2.1000 (-6.4180, 2.4430)*	0.1825 (-0.4751, 0.8419)
		IE, P	-2.5510 (-6.9740, 1.8710), P=0.26	2.5330 (-2.0150, 6.8540), P=0.26	0.1055 (-0.5990, 0.8089), P=0.77
	Regression coefficient for the interaction	**Direct**	0.1738 (-0.0974, 0.4451)	0.0126 (0.0006, 0.0245)	0.0191 (-0.0008, 0.0387)
		Indirect	0.0126 (0.0007, 0.0245)	0.1728 (-0.1048, 0.4376)	Fixed at zero
		IE, P	0.1613 (-0.1100, 0.4327), P=0.25	-0.1603 (-0.4253, 0.1173), P=0.24	0.0191 (-0.0008, 0.0387), P=0.06
Splitting the log odds ratio only (*model 2.2c*)	**Log odds ratio (centred)**	**Direct**	0.2495 (-0.3804, 0.8815)	0.4320 (0.2837, 0.5804)	0.2328 (-0.0031, 0.4700)
		Indirect	0.1994 (-0.0821, 0.4787)	0.4824 (-0.1946, 1.1600)	0.1816 (-0.4797, 0.8403)
		IE, P	0.0512 (-0.6481, 0.7515), P=0.89	-0.0499 (-0.7523, 0.6552), P=0.89	0.0521 (-0.6518, 0.7545), P=0.89
	Regression coefficient for the interaction	**All**	0.0129 (0.0011, 0.0248)	0.0129 (0.0011, 0.0248)	Fixed at zero
Splitting the regression coefficient only (*model 2.3c*)	**Log odds ratio (centred)**	**All**	0.1890 (-0.0918, 0.4673)	0.4283 (0.2793, 0.5747)	0.2746 (0.0469, 0.5033)
	Regression coefficient for the interaction	**Direct**	0.0195 (-0.0210, 0.0603)	0.0126 (0.0007, 0.0245)	0.0188 (-0.0007, 0.0385)
		Indirect	0.0125 (0.0007, 0.0245)	0.0194 (-0.0210, 0.0601)	Fixed at zero
		IE, P	0.0070 (-0.0358, 0.0500), P=0.75	-0.0068 (-0.0498, 0.0357), P=0.76	0.0188 (-0.0007, 0.0385), P=0.06

Table 3: Results from fixed-effect node-splitting models including common treatment by average age interactions for the malaria dataset.

AR: artemether; AS: artesunate; IE: inconsistency estimate; P: probability of agreement between direct and indirect evidence; QU: quinine.

*Results are influenced by the vague prior distribution and can be considered to be ‘not estimable’.

Model	Parameter	Posterior median (95% credibility interval)		
		AR vs. AS	**QU vs. AS**	**QU vs. AR**
NMR model (*model 1c*)	**Log odds ratio (centred)**	0.2080 (-0.0441, 0.4592)	0.4350 (0.2923, 0.5772)	0.2268 (0.0051, 0.4516)
	Regression coefficient for the interaction	0.0132 (0.0018, 0.0244)	0.0132 (0.0018, 0.0244)	Fixed at zero.
URM model assessing consistency of the log odds ratio and regression coefficient (*model 3.1c*)	**Log odds ratio (centred)**	0.2229 (-0.4006, 0.8471)	0.4365 (0.2891, 0.5832)	0.2743 (0.0363, 0.5136)
	Regression coefficient for the interaction	0.0145 (0.0044, 0.0247)	0.0145 (0.0044, 0.0247)	0.0145 (0.0044, 0.0247)
URM model assessing consistency of the log odds ratio only (*model 3.2c*)	**Log odds ratio (centred)**	0.2497 (-0.3819, 0.8806)	0.4317 (0.2831, 0.5794)	0.2328 (-0.0031, 0.4700)
	Regression coefficient for the interaction	0.0128 (0.0011, 0.0248)	0.0128 (0.0011, 0.0248)	Fixed at zero.
URM model assessing consistency of the regression coefficient only (*model 3.3c*)	**Log odds ratio (centred)**	0.1725 (-0.0811, 0.4257)	0.4402 (0.2978, 0.5822)	0.2676 (0.0416, 0.4959)
	Regression coefficient for the interaction	0.0148 (0.0048, 0.0246)	0.0148 (0.0048, 0.0246)	0.0148 (0.0048, 0.0246)

Table 4: Results from fixed-effect NMR and URM models with common treatment by average age interactions for the malaria dataset.
AR: artemether; AS: artesunate; NMR: network meta-regression; QU: quinine; URM: unrelated mean effects.

Dataset	Model	Mean residual deviance	p_D	DIC
Dataset 1: No interaction and consistency	**NMR model (*model 1a*)**	4.00	4.00	8.01
	Node-splitting model: AR vs. AS (*model 2.1a*)	6.00	6.00	12.00
	Node-splitting model: QU vs. AS (*model 2.1a*)	5.99	5.99	11.98
	Node-splitting model: QU vs. AR (*model 2.1a*)	5.99	5.99	11.98
	URM model (*model 3.1a*)	5.99	5.99	11.97
Dataset 2: Interaction and consistency	**NMR model (*model 1a*)**	4.00	4.00	8.00
	Node-splitting model: AR vs. AS (*model 2.1a*)	6.00	6.00	11.99
	Node-splitting model: QU vs. AS (*model 2.1a*)	5.99	5.99	11.99
	Node-splitting model: QU vs. AR (*model 2.1a*)	5.99	5.99	11.97
	URM model (*model 3.1a*)	5.98	5.98	11.97
Dataset 3: Interaction and inconsistency	**NMR model (*model 1a*)**	43.14	3.99	47.14
	Node-splitting model: AR vs. AS (*model 2.1a*)	5.99	5.99	11.99
	Node-splitting model: QU vs. AS (*model 2.1a*)	6.00	6.00	11.99
	Node-splitting model: QU vs. AR (*model 2.1a*)	5.98	5.98	11.97
	URM model (*model 3.1a*)	5.99	5.99	11.97
Dataset 4: No interaction and inconsistency	**NMR model (*model 1a*)**	184.36	4.00	188.36
	Node-splitting model: AR vs. AS (*model 2.1a*)	6.00	6.00	12.00
	Node-splitting model: QU vs. AS (*model 2.1a*)	5.99	5.99	11.99
	Node-splitting model: QU vs. AR (*model 2.1a*)	6.00	6.00	11.99
	URM model (*model 3.1a*)	5.99	5.99	11.98

Table 5: Model fit assessment results for fixed-effect models assessing consistency of both the log odds ratio and regression coefficient with independent treatment by average age interactions for the fabricated datasets.

Number of data points: 30

AR: artemether; AS: artesunate; DIC: deviance information criterion; QU: quinine; NMR: network meta-regression; URM: unrelated mean effects.

Dataset	Parameter	Evidence	Posterior median (95% credibility interval), P		
			AR vs. AS	QU vs. AS	QU vs. AR
Dataset 1: No interaction and consistency	Log odds ratio (uncentred)	Direct	0.1997 (-0.0948, 0.4949)	0.2302 (-0.0566, 0.5139)	0.0298 (-0.2356, 0.2937)
		Indirect	0.2001 (-0.1865, 0.5902)	0.2306 (-0.1642, 0.6265)	0.0297 (-0.3799, 0.4398)
		IE, P	-0.0007 (-0.4870, 0.4894), P=0.9974	-0.0004 (-0.4879, 0.4875), P=0.9986	-0.0002 (-0.4891, 0.4886), P=0.9990
	Regression coefficient for the interaction	Direct	0.0000 (-0.0107, 0.0109)	0.0000 (-0.0135, 0.0136)	0.0000 (-0.0115, 0.0116)
		Indirect	0.0000 (-0.0178, 0.0178)	0.0000 (-0.0158, 0.0158)	0.0000 (-0.0174, 0.0174)
		IE, P	0.0000 (-0.0210, 0.0208), P=0.9980	0.0000 (-0.0208, 0.0209), P=0.9980	0.0000 (-0.0208, 0.0209), P=0.9982
Dataset 2: Interaction and consistency	Log odds ratio (uncentred)	Direct	0.1992 (-0.1284, 0.5285)	0.2300 (-0.0268, 0.4852)	0.0301 (-0.3372, 0.3941)
		Indirect	0.1998 (-0.2432, 0.6460)	0.2304 (-0.2614, 0.7213)	0.0299 (-0.3886, 0.4447)
		IE, P	-0.0007 (-0.5528, 0.5534), P=0.9980	-0.0001 (-0.5549, 0.5537), P=0.9998	-0.0003 (-0.5542, 0.5548), P=0.9996
	Regression coefficient for the interaction	Direct	0.0200 (0.0049, 0.0352)	0.0200 (0.0069, 0.0333)	0.0000 (-0.0239, 0.0240)
		Indirect	0.0200 (-0.0073, 0.0473)	0.0199 (-0.0084, 0.0485)	0.0000 (-0.0200, 0.0201)
		IE, P	0.0000 (-0.0313, 0.0312), P=0.9974	0.0001 (-0.0315, 0.0313), P=0.9954	0.0000 (-0.0311, 0.0313), P=1.0000
Dataset 3: Interaction and inconsistency	Log odds ratio (uncentred)	Direct	0.2000 (-0.1389, 0.5372)	0.2301 (-0.0208, 0.4796)	0.0301 (-0.2355, 0.2937)
		Indirect	0.1999 (-0.1619, 0.5649)	0.2304 (-0.1985, 0.6584)	0.0299 (-0.3924, 0.4492)
		IE, P	0.0003 (-0.4955, 0.4950), P=0.9990	-0.0006 (-0.4948, 0.4955), P=0.9982	-0.0004 (-0.4971, 0.4983), P=0.9986
	Regression coefficient for the interaction	Direct	0.0100 (-0.0039, 0.0241)	0.0400 (0.0298, 0.0503)	0.0000 (-0.0125, 0.0126)
		Indirect	0.0400 (0.0237, 0.0562)	0.0099 (-0.0088, 0.0289)	0.0300 (0.0127, 0.0474)
		IE, P	-0.0300 (-0.0515, -0.0088), P=0.0059	0.0301 (0.0085, 0.0514), P=0.0062	-0.0300 (-0.0515, -0.0086), P=0.0057
Dataset 4: No interaction and inconsistency	Log odds ratio (uncentred)	Direct	0.2002 (-0.0926, 0.4908)	0.2300 (0.0222, 0.4360)	0.0297 (-0.2260, 0.2863)
		Indirect	0.2000 (-0.1290, 0.5298)	0.2300 (-0.1569, 0.6178)	0.0301 (-0.3279, 0.3866)
		IE, P	-0.0003 (-0.4376, 0.4397), P=0.9990	-0.0007 (-0.4393, 0.4399), P=0.9976	0.0000 (-0.4398, 0.4398), P=1.0000
	Regression coefficient for the interaction	Direct	-0.0400 (-0.0553, -0.0246)	0.0400 (0.0273, 0.0529)	0.0000 (-0.0115, 0.0116)
		Indirect	0.0399 (0.0227, 0.0574)	-0.0400 (-0.0591, -0.0208)	0.0800 (0.0600, 0.1000)
		IE, P	-0.0799 (-0.1031, -0.0571), P=0.0000	0.0800 (0.0568, 0.1030), P=0.0000	-0.0800 (-0.1031, -0.0569), P=0.0000

Table 6: Results from fixed-effect node-splitting models splitting both the log odds ratio and regression coefficient including independent treatment by average age interactions (*model 2.1a*) for the fabricated datasets.

Posterior median (95% credibility interval) presented.

AR: artemether; AS: artesunate; IE: inconsistency estimate; P: probability of agreement between direct and indirect evidence; QU: quinine.

Dataset	Model	Parameter	Posterior median (95% credibility interval)		
			AR vs. AS	**QU vs. AS**	**QU vs. AR**
Dataset 1: No interaction and consistency	**NMR model (*model 1a*)**	**Log odds ratio (uncentred)**	0.2002 (-0.0305, 0.4281)	0.2302 (0.0014, 0.4587)	0.0306 (-0.1911, 0.2517)
		Regression coefficient for the interaction	0.0000 (-0.0090, 0.0091)	0.0000 (-0.0102, 0.0102)	0.0000 (-0.0096, 0.0096)
	URM model (*model 3.1a)*	**Log odds ratio (uncentred)**	0.2002 (-0.0947, 0.4926)	0.2301 (-0.0556, 0.5148)	0.0303 (-0.2340, 0.2937)
		Regression coefficient for the interaction	0.0000 (-0.0108, 0.0108)	0.0000 (-0.0135, 0.0136)	0.0000 (-0.0116, 0.0116)
Dataset 2: Interaction and consistency	**NMR model (*model 1a*)**	**Log odds ratio (uncentred)**	0.2006 (-0.0539, 0.4514)	0.2302 (0.0043, 0.4558)	0.0298 (-0.2223, 0.2828)
		Regression coefficient for the interaction	0.0200 (0.0074, 0.0327)	0.0200 (0.0080, 0.0321)	0.0000 (-0.0147, 0.0147)
	URM model (*model 3.1a)*	**Log odds ratio (uncentred)**	0.2000 (-0.1289, 0.5266)	0.2301 (-0.0264, 0.4856)	0.0302 (-0.3364, 0.3948)
		Regression coefficient for the interaction	0.0200 (0.0049, 0.0351)	0.0200 (0.0068, 0.0332)	0.0000 (-0.0240, 0.0240)
Dataset 3: Interaction and inconsistency	**NMR model (*model 1a*)**	**Log odds ratio (uncentred)**	0.2081 (-0.0390, 0.4523)	0.1654 (-0.0503, 0.3808)	-0.0421 (-0.2636, 0.1801)
		Regression coefficient for the interaction	0.0187 (0.0082, 0.0292)	0.0335 (0.0244, 0.0425)	0.0147 (0.0047, 0.0248)
	URM model (*model 3.1a)*	**Log odds ratio (uncentred)**	0.2003 (-0.1374, 0.5353)	0.2301 (-0.0201, 0.4795)	0.0303 (-0.2340, 0.2938)
		Regression coefficient for the interaction	0.0100 (-0.0040, 0.0240)	0.0400 (0.0297, 0.0503)	0.0000 (-0.0125, 0.0125)
Dataset 4: No interaction and inconsistency	**NMR model (*model 1a*)**	**Log odds ratio (uncentred)**	0.0877 (-0.1296, 0.3034)	0.3389 (0.1566, 0.5214)	0.2515 (0.0472, 0.4567)
		Regression coefficient for the interaction	-0.0098 (-0.0211, 0.0017)	-0.0001 (-0.0105, 0.0103)	0.0097 (-0.0002, 0.0195)
	URM model (*model 3.1a)*	**Log odds ratio (uncentred)**	0.2004 (-0.0911, 0.4899)	0.2302 (0.0231, 0.4372)	0.0305 (-0.2259, 0.2854)
		Regression coefficient for the interaction	-0.0400 (-0.0553, -0.0247)	0.0400 (0.0272, 0.0529)	0.0000 (-0.0115, 0.0116)

Table 7: Results from fixed-effect NMR and URM models assessing consistency of both the log odds ratio and regression coefficient with independent treatment by average age interactions for the fabricated datasets.

AR: artemether; AS: artesunate; NMR: network meta-regression; QU: quinine; URM: unrelated mean effects.

Figure legends

Figure 1. Graphs showing how the relative treatment effect (e.g. log odds ratio) for treatment *3* vs. treatment *2* could change with a covariate value with separate lines representing direct evidence (from trials that allocated treatments *2* and *3*), indirect evidence (from the remaining trials), and all evidence in various scenarios: (a) there is no treatment by covariate interaction based on all evidence and the relative treatment effects at zero covariate are consistent and the regression coefficients for the treatment by covariate interaction are consistent; (b) there is an interaction based on all evidence and the relative treatment effects at zero covariate are consistent and the coefficients are consistent; (c) there is no interaction based on all evidence and the relative treatment effects at zero covariate are consistent and the coefficients are inconsistent; (d) there is an interaction based on all evidence and the relative treatment effects at zero covariate are consistent and the coefficients are inconsistent; (e) there is no interaction based on all evidence and the relative treatment effects at zero covariate are inconsistent and the coefficients are consistent; (f) there is an interaction based on all evidence and the relative treatment effects at zero covariate are inconsistent and the coefficients are consistent; (g) there is no interaction based on all evidence and the relative treatment effects at zero covariate are inconsistent and the coefficients are inconsistent; and (h) there is an interaction based on all evidence and the relative treatment effects at zero covariate are inconsistent and the coefficients are inconsistent.
Direct, indirect and all evidence is overlapping in plots (a) and (b).

Figure 2: Network diagram for the malaria dataset.
Number of trials (number of patients) displayed.
AR: artemether; AS: artesunate; QU: quinine.

Figure 3: Posterior distributions for the log odds ratios (centred) and regression coefficients for the interaction from fixed-effect node-splitting models with common treatment by average age interactions for the malaria dataset.
Results in figures a-f are from *models 2.1c* and *1c*. Results in figures g-i are from *models 2.2c* and *1c*. Results in figures j-l are from *models 2.3c* and *1c*. In figures f and i, the coefficient from indirect evidence and from all evidence is forced to be zero.
AR: artemether; AS: artesunate; QU: quinine.

Figure 4: Log odds ratio versus average age for direct and indirect from fixed-effect node-splitting models and for all evidence from the fixed-effect NMR model with common treatment by average age interactions for the malaria dataset.
Results in figures a-c are from *models 2.1c* and *1c*. Results in figures d-f are from *models 2.2c* and *1c*. Results in figures g-i are from *models 2.3c* and *1c*.

AR: artemether; AS: artesunate; QU: quinine.

Figure 5: Log odds ratio versus average age for direct and indirect from fixed-effect node-splitting models (*model 2.1a*) and for all evidence from the fixed-effect NMR model (*model 1a*) with independent treatment by average age interactions for the fabricated datasets.
AR: artemether; AS: artesunate; QU: quinine.