A framework for identifying treatment-covariate interactions in individual participant data network meta-analysis

S. C. Freeman[a, b], D. Fisher[a], J. F. Tierney[a], J. R. Carpenter[a,c]

a. MRC Clinical Trials Unit at UCL, Aviation House, 125 Kingsway, London, WC2B 6NH, UK

b. Department of Health Sciences, University of Leicester, University Road, Leicester, LE1 7RH, UK

c. London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

Abstract

Background: Stratified medicine seeks to identify patients most likely to respond to treatment. Individual participant data (IPD) network meta-analysis (NMA) models have greater power than individual trials to identify treatment-covariate interactions (TCI). TCI contain "within" and "across" trial interactions, where the across trial interaction is more susceptible to confounding and ecological bias.

Methods: We considered a network of IPD from 37 trials (5922 patients) for cervical cancer (2394 events), where previous research identified disease stage as a potential interaction covariate. We compare two models for NMA with TCI: (i) two effects separating within and across trial interactions and (ii) a single effect combining within and across trial interactions.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/jrsm.1300

We argue for a visual assessment of consistency of within and across trial interactions and consider more detailed aspects of interaction modelling, e.g. common vs trial-specific effects of the covariate. This leads us to propose a practical framework for IPD NMA with TCI.

Results: Following our framework, there was no evidence in the cervical cancer network for a treatment-stage interaction based on the within trial interaction. The NMA provided additional power for an across trial interaction over and above the pairwise evidence. Following our proposed framework, the within and across trial interactions should not be combined.

Conclusion: Across trial interactions are susceptible to confounding and ecological bias. It is important to separate the sources of evidence to check their consistency and identify which sources of evidence are driving the conclusion. Our framework provides practical guidance for researchers, reducing the risk of unduly optimistic interpretation of TCI.

Keywords: Framework; IPD; Network meta-analysis; Treatment-covariate interaction.

1. Background

Introduction

Stratified medicine seeks to identify patients most likely to respond to treatment. However, individual trials are rarely powered to detect interactions between treatment effects and participant characteristics. Meta-analysis (MA) models potentially have greater power to identify such treatment-covariate interactions (TCI), particularly when individual participant data (IPD) are available. One of the big advantages of IPD MA over aggregate data (AD) MA is the greater power that it affords for investigating treatment and patient-level covariate interactions (Berlin et al., 2002; Debray et al., 2015; Lambert et al., 2002; Tierney et al., 2015). To explore how the treatment effect may vary in relation to a patient-level covariate, a TCI can be fitted (Donegan et al., 2012; Hua et al., 2017). The power to detect TCI will depend on the distribution of the covariate within each study (Simmonds and Higgins, 2007).

Pairwise IPD MA of TCI often results in two sources of information; so-called 'within trial' (at the individual patient level) and 'across trial' (at the trial level) interactions, where the across trial interaction is particularly susceptible to confounding and ecological bias as it is based on observational associations (Berlin et al., 2002; Debray et al., 2015; Hua et al., 2017; Riley et al., 2008; Riley and Steyerberg, 2010; Simmonds and Higgins, 2007; Thompson and Higgins, 2005).The same is true of IPD network meta-analysis (NMA); therefore it is important that the within and across trial interaction estimates are reviewed separately, before deciding whether to combine them. We reiterate that confounding from unmeasured covariates (e.g. differences in baseline disease risk) can affect both within and across trial interactions (Jackson et al., 2006). However, within trial effects are less susceptible because of the protection provided by the trials randomisation. Ecological bias arises if across trial information, which uses trial average covariate values, isused to draw conclusions about individuals (Greenland, 1987; Hong et al., 2015; Morgenstern, 1982; Simmonds and Higgins, 2007). However, if interest lies in the population level effect and it is interpreted correctly as a population effect (and not an individual effect) then (by definition) there cannot be ecological bias. In some instances, it may be the case that the across trial interaction (e.g. the mean effect for a patient aged 50) truly differs from the within trial interaction (e.g. the effect for an individual patient aged 50) (Simmonds and Higgins, 2007). Separating out within and across trial interactions may change the conclusions drawn from combining within and across trial interactions. For example, an IPD MA comparing two anti-epileptic drugs as monotherapy for controlling seizures originally combined the within and across trial interactions and identified an interaction between treatment and age (Tudur Smith et al., 2005). However, a recent re-analysis separating out the within and across trial interactions no longer indicated an interaction between treatment and age based on the within trial interaction only (Hua et al., 2017).

IPD MA models separating out within and across trial interactions were first developed for continuous outcomes (Riley et al., 2008) and later applied to time-to-event outcomes (Fisher et al., 2011). In the IPD NMA setting models separating out within and across trial interactions have been proposed for dichotomous outcomes (Donegan et al., 2012) and time-to-event outcomes using the Cox regression model (Hua et al., 2017). In this paper we show how to separate out the within and across trial interactions in the IPD NMA setting for time-to-event outcomes using the Royston-Parmar model.

A key assumption of NMA is consistency between the direct and the indirect evidence (Salanti, 2012; Dias et al., 2013). The inclusion of TCI in a NMA model offers one of many ways for exploring and understanding inconsistency (Cooper et al., 2009; Donegan et al., 2012, 2013a,b). With the presence of TCI in a NMA model, the consistency assumption may be violated if one or more of the true treatment effects is modified by a covariate and included trials differ with respect to the covariate (Donegan et al., 2012). Furthermore, when we include TCI in a NMA model we assume that the treatment effects estimated at the covariate value of zero are consistent and that the regression coefficients for the TCI parameters are also consistent (Donegan et al., 2017). Therefore, it is important to assess the consistency assumption either at each level of the covariate (for categorical covariates) or across a range of values (for continuous covariates) (Donegan et al., 2017).

There are three different ways to model TCI effects: common, independent or exchangeable (Cooper et al., 2009). Common effects assume that the regression coefficients are the same for all TCI so that the TCI effect is the same for each treatment compared tothe control. Independent effects assume that all TCI are different for each treatment versus the control so that a separate regression coefficient for each TCI is included in the model. Exchangeable effects assume that all TCI are different from each other but similar enough that they can be sampled from a common distribution. For IPD MA, and as we show by extension in IPD NMA, there are three possible ways of analysing TCI: using the across trial interaction only, using the within trial interaction only and combining the two (Fisher et al., 2017). We now describe three approaches for IPD MA with TCI before considering the NMA setting.

In one commonly used approach for interactions in MA (which can be shown to combine within and across trial interactions), specifically for categorical covariates, the treatment effect is calculated within each trial for each level of the covariate. The treatment effects for each level of the covariate are combined across all trials, using standard MA techniques, resulting in an overall effect for each level of the covariate which are then compared to each other (Fisher et al., 2011, 2017; Riley et al., 2008). Any trials where all patients have the same covariate value will not contribute to the within trial interaction but can contribute to the across trial interaction. This is a common approach used in IPD MA. However, the within trial interaction can be exaggerated or masked by the across trial interaction, which is at risk of ecological bias. Therefore, this approach is also at risk of ecological bias (Fisher et al., 2011, 2017; Riley et al., 2008).

An analysis using the across trial interaction only considers how the treatment effect varies across trials in relation to the trial mean value of the covariate and fails to use the patient-level information (Thompson and Higgins, 2002). This requires the assumption of no unmeasured confounding between the outcome and the covariate, and that there is no ecological bias (Fisher et al., 2017; Riley et al., 2008; Riley and Steyerberg, 2010). Unfortunately it is typically not possible to identify such confounders as baseline data often varies across trials. Therefore it is often not possible to test whether the inclusion of the across trial interaction will induce bias.

An analysis using the within trial interaction only more closely parallels the underlying principles of MA. Estimates of the TCI effect are calculated within each trial and then pooled together using MA methods (Simmonds and Higgins, 2007; Riley et al., 2008; Fisher et al., 2011). Any trials where all patients have the same covariate value will not contribute to this analysis as they do not provide any within trial interactions (Fisher et al., 2017; Riley et al., 2008; Simmonds and Higgins, 2007). Recommendations on the presentation and analysis of TCI using this approach are proposed by Fisher (2017). A key aim of this paper is to show how these recommendations can be brought to bear in NMA.

In the pairwise MA case, it is clear that within trial interactions are the most clinically relevant estimates, as they are free from ecological bias (Berlin et al., 2002; Donegan et al., 2012; Fisher et al., 2017; Hua et al., 2017; Riley et al., 2008). In the NMA case, more research is needed to explore how the consistency assumptions of the network applied to within trial interactions (as in our model) can help to improve their precision. Nevertheless, the framework described here explicitly separates the within and across trial interactions throughout the network, and hence guarantees that an unbiased estimate of the within trial interaction is obtained. The methods proposed here are applicable to both continuous and categorical covariates. Specifically, in this paper we illustrate how the within and across trial interactions can be separated for time-to-event outcomes modelled using the IPD Royston-Parmar NMA model, we propose a framework for conducting NMA with TCI, showing how to fit models which separate out the within and across trial interactions. We then illustrate our framework by applying it to a cervical cancer network.

Why do we need a framework?

Fitting a NMA with TCI is often a more complex process than researchers anticipate. There are a number of additional important decisions that need to be taken, beyond those that need to be considered in a MA. These include the parameterisation and consistency of covariate and interaction effects. An added complication, frequently encountered in practice, is how to handle missing patient-level covariate data.

When it comes to reporting a NMA with TCI, the preferred reporting items for systematic reviews and meta-analyses based on IPD (PRISMA-IPD) guidelines recommend pre-specification of whether the across trial interaction is to be combined with the within trial interaction (Stewart et al., 2015; Tierney et al., 2015). Therefore, it is important that authors are aware of the potential implications of combining within and across trial interactions. By proposing a framework for conducting NMA with TCI we aim to equip researchers with the knowledge and tools for successfully fitting an appropriate NMA with TCI.

In Section 2 we discuss issues to consider before conducting NMA with TCI, outline a nine-step framework for one-stage IPD NMA with TCI, provide guidance on implementing the framework and introduce a cervical cancer dataset. In Section3 we present the results of applying the framework to the cervical cancer network. In Section 4 we discuss the framework before drawing some conclusions in Section 5.

2. Methods

Issues to consider

Preliminary analysis

NMA often starts with a systematic review being conducted to identify all treatments and trials to be considered in the network. As part of the review and in discussion with appropriate clinicians, discussion of any covariates which could be included in a NMA with TCI should take place before any models are fitted. Such models require a number of considerations and preliminary analysis of the data can help inform the decision of which model to fit.

Common main effect vs trial-level main effect of covariate

A patient-level covariate can be fitted as a common effect or a trial-level effect (Govan et al., 2010). A common effect pools the effect of the covariate across all trials. A fixed trial effect results in a separate estimate of the effect of the covariate for each trial and does not provide an overall effect for the covariate. A random trial effect of a covariate allows the effect of the covariate to differ in each trial assuming that the coefficients for each trial come from a common (typically normal) distribution.

We encourage the use of a trial-level effect, either fixed or random. If a common effect of a covariate is used when a trial-level effect would be more appropriate this can result in a poorly fitting model which could affect convergence, suppress the differences between trials and affect the treatment effect estimates. Assuming a common effect of a covariate is generally not appropriate when the distribution of the covariate varies between trials or in a network where trials vary in size, because it is known that smaller studies can give more extreme parameter estimates (Chaimani and Salanti, 2012). A hypothesis test to check the effect of the covariate in each trial should be conducted before assuming a common effect, as this choice is likely to critically impact the estimate of the TCI.

Parameterisation of within and across trial interactions

TCI can be included in MA models in two ways. Firstly, as a single effect which combines within and across trial interactions; and secondly, as two effects which separate out the within and across trial interactions (Riley et al., 2008; Fisher et al., 2011). We now describe how to do this in the NMA setting using the Royston-Parmar model for time-to-event data.

Consider the one-step fixed treatment effect Royston-Parmar NMA model for a network of $q+1$ treatments (Freeman & Carpenter, 2017). Including a fixed trial effect of a patient-level covariate z_{ij}, the log cumulative hazard for patient *i* from trial *j* can be modelled as:

$$\ln\left\{H_j\left(t \mid X_{ij}\right)\right\} = s_j\left(\ln\left(t\right)\right) + \beta_1 \mathrm{trt1}_{ij} + \cdots + \beta_q \mathrm{trt}_{ij} + a_j z_{ij} \qquad (1)$$

where $s_j(\ln(t))$ is the restricted cubic spline for trial j, $\mathrm{trt1}_{ij}, \ldots, \mathrm{trtq}_{ij}$ are the treatment indicators with corresponding coefficients $\beta_1, \ldots, \beta_q$ and α_j is the effect of the patient-level covariate z_{ij} for trial *j*.

Adding a common TCI to (?) which separates the within and across trial interactions results in:

$$\begin{aligned} \ln\left\{H_j\left(t \mid x_{ij}\right)\right\} &= s_j\left(\ln\left(t\right)\right) + \beta_1 \mathrm{trt1}_{ij} + \cdots + \beta_q \mathrm{trtq}_{ij} + a_j\left(\bar{z}_j\right) \\ &+ \delta_{A1} \mathrm{trt1}_{ij}\left(z_{ij} - \bar{z}_j\right) + \cdots + \delta_{Aq} \mathrm{trtq}_{ij}\left[z_{ij} - \bar{z}_j\right] \\ &+ \delta_{B1} \mathrm{trt1}_{ij} \bar{z}_j + \cdots + \delta_{Bq} \mathrm{trtq}_{ij} \bar{z}_j \end{aligned} \qquad (2)$$

where the covariate z_{ij} is fitted as a fixed trial-level effect with coefficient α_j for trial *j* and $\bar{z}_j$ is the mean value of z_{ij} for trial *j*. The within trial interaction is represented by the $a_{A1}, \ldots, \delta_{Aq}$ parameters whilst the across trial interaction is represented by the $\delta_{B1}, \ldots, \delta_{Bq}$ parameters. The difference $\delta_{Bk} - \delta_{Ak}$ quantifies the amount of ecological bias for interaction *k* (Fisher et al., 2011).

Adding a common TCI to (1) which combines the within and across trial interactions results in:

$$\ln\{H_j(t|x_{ij})\} = s_j(\ln(t)) + \beta_1 \mathrm{trt1}_{ij} + \cdots + \beta_q \mathrm{trtq}_{ij} + \alpha_j z_{ij}$$

$$+\delta_1 \mathrm{trt1}_{ij} z_{ij} + \cdots + \delta_q \mathrm{trtq}_{ij} z_{ij} \qquad (3)$$

where the covariate z_{ij} is fitted as a fixed trial-level effect with coefficient α_j for trial j and $\delta_1, \dots, \delta_q$ are the coefficients of the TCI effects.

In practice, these models can also be fitted with random treatment effects, random trial-level treatment-covariate interactions and random trial-level effect of the covariate (Appendix A).

Missing covariate data

NMA can be conducted in both the frequentist and Bayesian frameworks (Lu and Ades, 2004; Lumley, 2002). One of the advantages of conducting NMA within a Bayesian framework, and in particular using WinBUGS, is that missing covariate data can be naturally handled in WinBUGS. Missing covariate data can be accommodated within the NMA model by including a distribution for the covariate with missing values. This allows two things to happen: missing covariate values are imputed which allows a patient to be included in the NMA model and this in turn increases the precision of the treatment effects which themselves inform the imputation of the missing values (Lunn et al., 2013). If we wish to perform a frequentist analysis, the most straightforward way to handle missing covariate values is by multiple imputation; however, this is not a straightforward application of multiple imputation, because the imputation needs to be done in a way that is consistent with the NMA model. The R-pacakge jomo (Quartagno and Carpenter, Stat Med, 2016) has the flexibility to handle this, but we do not pursue this further.

Nine-step framework for one-stage IPD NMA with treatment-covariate interactions

The aim of this framework is to provide guidance on the steps that need to be considered before a NMA with TCI can be fitted, so that the analysis is conducted systematically and appropriately. This framework concentrates specifically on building NMA with TCI models. Therefore, the framework assumes that the usual MA activities such as protocol writing, defining inclusion/exclusion criteria and determining whether trials are similar enough for inclusion in a MA have already been conducted. Further, we assume that the network is connected and any covariates for inclusion in the NMA models have been identified through discussion with clinicians. As usual, when interpreting the results, a range of possible causes of heterogeneity (e.g. baseline differences, design, treatment doses, delivery, escape therapies) must be kept in mind. Steps 1-4 are applicable for any NMA whether covariates are considered for inclusion or not. From step 5 onwards the framework specifically considers the inclusion of covariates and TCI. This framework has been developed to be applicable to a range of outcomes (e.g. binary, continuous and time-to-event). As is common, we work on the log-odds or log-hazard scale. However, the principles underlying our proposed framework will apply to other settings but the technical details may vary. The framework may need to be tweaked to take into account additional issues arising in specific settings.

1. Assess all pairwise treatment comparisons for evidence of heterogeneity
   (a) If heterogeneity is present, explore the baseline characteristics of all trials. Can the heterogeneity be explained by differences in baseline characteristics across trials?
      i. If yes, all important covariates should be considered going forwards
      ii. If no, it could be unsuitable to combine the pairwise comparison in a NMA
2. Identify a reference treatment across the network and determine which treatment contrasts will be parameterised in the model (all other treatment contrasts should be obtained through consistency equations)

3. Fit the NMA model without covariates taking into account any heterogeneity identified in step 1 (e.g. by using a random effect model)
4. Assess the network for evidence of inconsistency
5. Investigate patterns of missing data for the covariate of interest
6. Consider modelling assumptions for including the covariate in the NMA model (fixed or random treatment effects? common, fixed-trial or random-trial effect of covariate?)
7. Fit NMA model including covariate and assess model results
8. Fit NMA model including TCI with within and across trial interactions separated and assess agreement between the within and across trial interactions
9. Fit NMA model including TCI with within and across trial interactions combined, if appropriate

Note, when there is missing covariate data it is often practically and computationally easier to fit step 9 before step 8. More details can be found in the guidance on implementing the framework section.

Guidance on implementing the framework

Step 1: Before a NMA model is fitted all pairwise treatment comparisons in the network should be explored for evidence of heterogeneity. Heterogeneity can be assessed through the I^2, τ^2 and Cochran's Q statistics (Cochran, 1954; Higgins et al., 2003; R¨ucker et al., 2008). If heterogeneity is present, explore the baseline characteristics of all trials. If one trial, or a subgroup of trials, are found to be causing the heterogeneity then exploring the baseline characteristics can identify what isdifferent about this trial, or trials, and the impact this might have on the treatment effect. The identification of heterogeneity, at this stage, in one or more pairwise comparisons can determine whether fixed treatment effect (FTE) or random treatment effect (RTE) models are used in step 4 (Mills et al., 2013). If the source of heterogeneity cannot be identified or accounted for then either RTE will need to be used or it will be unsuitable to use this comparison in a NMA, particularly if removingthe pairwise comparison exhibiting heterogeneity means that a FTE model can be used. Any covariates identified, during this step, as potentially causing heterogeneity should be considered in steps 6 to 9.

Step 2: The previous standard of care, the largest treatment node or the treatment connected to the greatest number of other treatments are the most appropriate choices for the reference treatment within the network. The treatment parameterisation of the network should satisfy the consistency equations (Higgins and Whitehead, 1996). The number of treatment parameters should be one less than the number of treatments in the network. The network diagram can help inform which treatment parameters should be directly estimated in the NMA model and which will be calculated as contrasts through the consistency equations. Network diagrams can be created in Stata [52] using the *networkplot* command (Chaimani et al., 2013).

Step 3: A NMA model can be fitted using both FTE and RTE and monitoring the deviance information criteria (DIC). If heterogeneity was present in step 2, the RTE model should be used as this increases the variability around the point estimate to reflect the heterogeneity. The RTE model gives more weight to smaller studies than the FTE model. Therefore, a difference in the treatment effect estimates between the FTE and RTE models can indicate publication bias and small study effects (Chaimani and Salanti, 2012). DIC is a measure of model fit which penalises model complexity - smaller values are better. DIC can be used to compare models although small differences (i.e. < 5) should not be over-interpreted and simpler models should be chosen where they can be (Spiegelhalter et al., 2002). In addition, in some cases, total residual deviance can be used to assess goodness of fit; see Dias (2013) for details.

Step 4: The network should be assessed for evidence of inconsistency (Caldwell et al., 2005; Donegan et al., 2012, 2013b; Ioannidis, 2009; Lu and Ades, 2006; Salanti et al., 2007; Veroniki et al., 2013). To visualise this, it is useful to present the model results as a forestplot with the network, direct and indirect evidence separated out. There are many approaches to assessing inconsistency (e.g. node-splitting (e.g. node-splitting (Dias et al., 2010; van Valkenhoef et al., 2016; Tu, 2016), inconsistency models such as the design-by-treatment interactionmodel (Higgins et al., 2012; White et al., 2012), random inconsistency effects (Jackson et al., 2014, 2016; Law et al., 2016), factorial analysis of variance (Piepho, 2014), generalised linear mixed models (Tu, 2015; Günhan et al., 2017)) and the two-stage approach (Lu et al., 2011)). We recommend consulting review papers such as Donegan et al (2013a) and Efthimiou et al (2016) which describe and compare different methods for assessing consistency to help select the most appropriate method for the network at hand.

We use the inconsistency parameter approach of Lu & Ades (2006) in which an inconsistency parameter is fitted for each treatment loop and the model is re-fitted including the additional parameters. This approach complements the assessment of heterogeneity from step 1 and follows the approach outlined in Freeman & Carpenter (2017). Here an inconsistency parameter is initially added to the FTE model before considering the RTE model and exploring whether the conclusions are sensitive to the inclusion of the inconsistency parameter. If inconsistency is present in the network then an inconsistency parameter can be used in all further models. Treatment loops with inconsistency parameters are reduced to thedirect evidence only and therefore do not contribute to the across trial interaction in the network. Furthermore, if inconsistency is present, the cause of the underlying inconsistency/heterogeneity must be resolved before the results are used for clinical inference (Donegan et al., 2017; Freeman and Carpenter, 2017).

Step 5: Consider the distribution of the covariate of interest in each trial. Are there any trials where some patients have missing covariate data? Are there any trials where all patients have missing covariate data? Is the covariate continuousor categorical? Can a linear effect between the groups of an ordered categorical covariate be assumed? What is the reference value of the covariate? In WinBUGS (Lunn et al., 2000), as in all Bayesian modelling software, missing covariate data can be imputed once the marginal distribution of the covariate is specified.

Step 6: Covariates can be included as common effects, fixed trial-level effects or random trial-level effects. DIC can be used to determine which of these assumptions is most appropriate. However, assuming a common effect of a covariate is only likely to be appropriate if the distribution of the covariate is the same in every trial in the network. The choice of FTE or RTE should be informed by previous steps such as the presence of heterogeneity from step 2 or the DIC from step 4.

Step 7: Fit the NMA models including the patient-level covariate and assess the results. This can help inform the decision of which NMA model with TCI to fit. TCI can be fitted as either common effects, fixed trial-level effects or random trial-level effects. A common effect assumes that the TCI has the same effect in all trials. A random trial-level effect allows the effect of the TCI to differ in each trial but assumes that the coefficients for each trial come from a common (typically normal) distribution. The choice of assumption for TCI can be informed by the distribution of the covariate within and across trials.

Steps 8 & 9: The framework recommends that the within and across trial interactions are considered separately at first and then combined if it is appropriate to do so. However, the mean covariate value from each trial is needed to separate out the within and across trial interactions and to calculate this missing covariate data needs to be imputed. Although it is possible, in principle, to impute the missing covariate data, calculate the mean covariate value and fit the NMA model separating within and across trial interactions in one step, in practice software is unlikely do this. Therefore it is more practical and computationally easier to fit the model combining within and across trial interactions first and monitor the mean value of the covariate in any trials with missing covariate data before using these values to fit the model separating the within and across trial interactions. For trials with only some missing covariate data the weighted average of the mean observed value and mean imputed value of the covariate can be used as the trial mean value. In trials where all patients have missing covariate values the mean covariate value from the imputed values can be used as the trial mean value. A sensitivity analysis in which patients with missing covariate data are excluded can be conducted to check the imputation of the missing covariate data has been handled correctly.

A visual assessment of the agreement between the within and across trial interactions can be made by plotting the parameter estimates for the TCI. Log hazard ratios along with 95% credible intervals can be presented in tables. If TCI are present, the treatment effect parameters on their own do not have a useful interpretation. Treatment effects should be presented separately for each level of the covariate. Consistency can then be checked for each level of the covariate following methods described by Donegan et al (2017). Graphs ranking the treatments for each level of the covariate can be used as a visual aid for determining the most effective treatment for each level of the covariate (Chaimani et al., 2013; Salanti et al., 2011).

Example

Our example, comes from three meta-analyses of randomised controlled trials (RCTs) in cervical cancer performed by two international collaborations (Chemoradiotherapy for Cervical Cancer Meta-Analysis Collaboration, 2008; Neoadjuvant Chemotherapy for Cervical Cancer Meta-analysis Collaboration, 2003). The three meta-analyses considered four different treatments: radiotherapy (RT), chemoradiation (CTRT), neoadjuvant chemotherapy plus radiotherapy (CT+RT) and neoadjuvant chemotherapy plus surgery (CT+S) (Figure 1).

The RT v CTRT comparison included a total of 18 RCTs and 4818 patients. In the original publication five of these trials were only included in sensitivity analyses as patients on at least one of the treatment arms received additional treatment. This resulted in a subset of 13 trials (3104 patients) which were identified and used for the primary analysis. Within this subset of 13 trials one three-arm trial combined two different forms of CTRT and compared them with a single control arm and three four-arm trials were split into two unconfounded comparisons of RT v CTRT for analysis as separate trials. This resulted in 16 trials included in the primary analysis. In this paper we will only consider the trials used in the primary analysis and will treat the data in the same way as the original publication (Chemoradiotherapy for Cervical Cancer Meta-Analysis Collaboration, 2008).

Across the three meta-analyses that form our network of trials, overall survival data was available for 5922 patients from 37 RCTs (35 two-arm RCTs, 2 three-arm RCTs). Covariate data was available for stage of disease from 5517 patients from 36 RCTs.

3. Results

Application of framework to cervical cancer network

In this section we illustrate the application of the proposed framework for one-stage IPD NMA with TCI to the cervical cancer network. In this example we use the one-stage IPD Royston-Parmar NMA model in the Bayesian setting to analyse overall survival (Freeman and Carpenter, 2017). Based on the availability of IPD stage of disease will be considered for inclusion in a NMA model with TCI. Based on a test of linearity we treat stage of disease as linear throughout the rest of this paper. All models were fitted in WinBUGS version 1.4.3 (Lunn et al., 2000) and run with 20,000 burn-in and 20,000 iterations and two sets of initial values. Convergence was checked by examining the trace and histograms of the posterior distribution. Models were compared using the DIC statistic (Lunn et al., 2013; Spiegelhalter et al., 2002). Parameters representing the spline function for the baseline log cumulative hazard function, treatment effects and inconsistency parameters were fitted with non-informative normal prior distributions. In the random treatment effect (RTE) model the treatment effects were modelled using a multivariate normal distribution with the mean coming from a normal distribution and precision from a Wishart distribution. Parameter estimates are presented as log hazard ratios (LogHR) and 95% credible intervals (CrI) for the posterior mean. A LogHR of zero indicates a null effect and a LogHR less than zero indicates a beneficial effect relative to the reference treatment, RT.

Step 1: All pairwise treatment comparisons were assessed for evidence of heterogeneity using Cochran's Q statistic and the I^2 statistic (Cochran, 1954; Higgins et al., 2003). There was no evidence of heterogeneity within the RT v CTRT (p=0.625, Table 1) and CT+RT v CT+S (p=0.939) comparisons while there was some evidence of statistical heterogeneity in the RT v CT+S (p=0.065) comparison and substantial heterogeneity in the RT v CT+RT comparison ($p<0.001$, also noted in the original publication (Neoadjuvant Chemotherapy for Cervical Cancer Meta-analysis Collaboration, 2003)). The baseline characteristics of all trials were compared. A pre-specified analysis of RT v CT+RT identified a difference in treatment effect by chemotherapy cycle length. Therefore, CT+RT was split into two treatments based on the length of chemotherapy cycles. Throughout the rest of this paper trials with chemotherapy cycles less than or equal to 14 days will be referred to as 'short cycles' and trials with chemotherapy cycles greater than 14 days will be referred to as 'long cycles'. No evidence of heterogeneity was found in the RT v CT+RT long cycles comparison (p=0.263). However, there was evidence of heterogeneity in the RT v CT+RT short cycles comparison (p=0.002). Heterogeneity can also be assessed visually from the forest plots in Figure 2. Treatment effects are presented in Table 1.

We also assessed the assumption of proportional hazards (PH). Following the methods described by Freeman and Carpenter (2017), we performed a global test of non-proportionality which was not significant for any of the pairwise comparisons; therefore we continue under the assumption of PH in the cervical cancer network.

Step 2: RT was chosen as the reference treatment because it was the previous standard of care. We included the RT v CTRT, RT v CT+RT and RT v CT+S treatment contrasts as parameters in the NMA model resulting in the treatment effect for CT+RT v CT+S being estimated through the consistency equation.

Step 3: A one-stage IPD Royston-Parmar NMA model was fitted to the cervical cancer network using both FTE and RTE with the results presented in Table 2. The DIC provides only weak evidence in favour of the RTE model (FTEDIC=12321.5, RTE DIC=12315.8). However, due to the presence of heterogeneity in the RT v CT+S short cycles comparison, identified in step 1, the RTE model was deemed to be the most appropriate model.

Step 4: In Figure 3 the direct and indirect treatment effects differ from each other with the network estimates balancing out these two sources of information. The direct and indirect treatment effects are estimated through the inclusion of an inconsistency parameter which was estimated as -0.484 (95% CrI: -1.314, 0.354). Cochran's Q statistic showed some evidence of inconsistency between designs (Q=10.32, 2 df, p=0.006). The inconsistency between designs is driven by one trial (Sardi et al., 1996) which had a treatment effect estimate more extreme than the other trials.

In addition, we also assessed globally the assumption of proportional hazards across the network using the method recommended by Freeman and Carpenter (2017). The Wald test for non-PH from the RTE model with random treatment-ln(time)interactions gave $\chi^2 = 0.324$ on 3 degrees of freedom (p=0.955) giving no evidence of non-PH within the network.

Step 5: A linear effect of stage of disease was assumed which could take the values 0 = stages 1A-2A, 1 = stage 2B, 2 = stages 3A-4A. We assessed this assumption by conducting a Wald test for each trial which included patients covering all three categories of stage of disease. As each trial is independent, we summed together the chi-squared statistics to provide an overall test of the linearity assumption. This gave χ^2=8.19 on 12 degrees of freedom and p=0.77. Therefore, we proceeded withthe assumption of a linear effect for stage of disease.

Thirteen trials had at least one patient with missing stage data. One of these trials had missing stage data for all patients. To impute values of missing stage of disease (whether explicitly, using multiple imputation, or as part of a Bayesian model) we need to assume a distribution. We used a truncated normal distribution which we believe is a reasonable approximation for a clinical severity measure of this kind. This is especially so as (in common with other settings) the majority of information is recovered by bringing the observed data on patients with missing stage into the model, and not on the stage coefficients themselves. In WinBUGS, the normal distribution was truncated through the use of the 'I' function to restrict missing covariates to take values between 0 and 2. Results with a categorical stage model were similar, but we found it harder to obtain convergence

Step 6: A common effect of stage of disease appeared to be inappropriate as the distribution of stage of disease varies across trials and the network includes trials of varying sizes. In addition, the DIC showed that a random effect of stage was most appropriate. Therefore, a RTE model with random trial-level effect of stage of disease was fitted.

Step 7: As expected, when included as a covariate, the parameter estimate for stage of disease suggests that overall survival is reduced as stage of disease increases (LogHR=0.561, 95% CrI: 0.475, 0.641; Table 2). Despite the inclusion of stage of disease as a covariate, the treatment effect for CTRT compared to RT remained statistically significant.

Step 8: A RTE model with fixed trial-level effect of stage of disease and random trial-level effect of treatment-stage interactions separating out the within and across trial interactions was fitted. There were no statistically significant interactions between treatment and stage of disease (Table 3). The credible intervals for the RT v CT+S comparison are much wider, relative to the other treatment comparisons, possibly reflecting the small amount of within trial interaction. In this comparison, there are only two trials which have patients distributed over more than one value of stage and can therefore contribute to the within trial interaction.

A visual assessment of the consistency of the within and across trial interactions was conducted by plotting the parameter estimates for the treatment-stage interactions (Figure 4). To determine whether any information was gained from the NMA we also plotted the MA estimates from a FTE model (Figure 4). We used a relatively strict criterion considering agreement to be shown if the within trial interaction was within half a standard error of the across trial interaction.For the RT v CTRT comparison there is agreement between the within and across trial interactions. This is in line with our expectations as the RT v CTRT comparison was a branch of the network without any indirect evidence informing the comparison. However, in the case of RT v CT+RT and RT v CT+S the within and across trial interactions do not agree. For example, the LogHR of the within trial interaction for CT+RT is -0.035 (95% CrI: -0.285, 0.204) and the LogHR for the across trial interaction is 0.110 (95% CrI: -0.315, 0.545; Table 3). The within and across trial interactions are not consistent with each other and the across trial interaction could be subject to ecological bias. Therefore we should focus on the within trial interaction only. For the RT v CT+RT and RT v CT+S comparisons further investigation into the difference between the within and across trial interactions may be required as these comparisons could be subject to ecological bias.

Step 9: To fully illustrate our framework we also fitted a model combining the within and across trial interactions (Table 3). However, as mentioned in step 8, the within and across trial interactions should remain separated. Additionally, a sensitivity analysis in which patients with missing stage of disease were excluded was conducted (Table B.1, Appendix B). Table 3 and Table S.1 show good agreement between the two models. However, in Table S.1 the across trial interaction from the CT+S and stage interaction was statistically significant.

4. Discussion

NMA with TCI has the potential to identify groups of patients most likely to respond to treatment. We have proposed a practical framework which aims to encourage researchers to conduct one-stage IPD NMA with TCI in a systematic and appropriate manner. We have successfully applied the framework to a cervical cancer network. The framework highlights the importance of preliminary analyses to consider issues such as heterogeneity and inconsistency which may inform decisions around the most appropriate modelling assumptions. This framework is deliberately generic, so that it can be applied to a range of outcomes (e.g. binary, continuous and time-to-event) and has the potential to improve the conduct and analysis of NMA with TCI.

In the cervical cancer network we showed that stage of disease had a statistically significant effect on overall survival with advanced disease increasing the risk of death. Due to the presence of heterogeneity in the network a RTE model was considered to be the most appropriate. There was no evidence of a treatment-stage interaction based on the within trial interaction for any of the treatments leading to the conclusion that stage of disease did not modify the treatment effect. Based on our relatively strict criterion, the treatment-stage interaction models showed a difference between the within and across trial interactions for some of the comparisons and it was therefore most appropriate to separate out the within and across trial interactions. The small difference in the NMA and MA estimates of the across trial interaction suggested that some across trial interaction might have been gained from the network. Our criterion for assessing agreement between the within and across trial interactions is arguably somewhat strict and arbitrary. However, we feel it is better to be cautious as a number of treatment-covariate interactions identified in the literature have subsequently been debunked (Hua et al., 2017).

The cervical cancer network is a small, well-connected network with a lot of direct evidence. Despite this we were still able to show that some across trial interaction is gained when conducting a NMA. Information (in the statistical sense of the inverse of the squared standard error) is gained in a NMA when the direct and indirect evidence are consistent. In practice, not all networks will contain as much direct evidence as the cervical cancer network. Therefore, we would expect NMA to contribute agreater amount of statistical information on an across trial interaction in a consistent network where some treatment comparisons are only informed by a small amount of direct evidence. We may also expect to gain more information when using a fixed effects NMA.

A reviewer suggested that (as the IPD data are available) the following simpler analysis may be preferable: (i) perform a two-stage MA of the within-study interaction effects, then (ii) derive the interaction effects within each study using the IPD and then perform NMA on these. In applications, this could provide a useful cross-check of the results. However, in our setting the network gives us the best treatment estimates against which to estimate an interaction. Further, NMA goes beyond just estimating treatment and interaction effects: a key motivation for NMA is to rank treatments in terms of efficacy. The inclusion of treatment-covariate interactions allows us to consider whether the ranking of treatments varies by covariate level. Indeed, if a treatment-covariate interaction is present, we should rank treatments for each level of the covariate.

With the cervical cancer network we did not need to include an inconsistency parameter. However, in a network with one treatment loop if an inconsistency parameter is included then it may also be appropriate to allow for inconsistency in the TCI. This would be equivalent to conducting separate pairwise MA with TCI and nothing would be gained from conducting a NMA. We also only considered one covariate whereas in practice researchers may wish to consider multiple covariates. In this case, we would recommend considering each covariate on its own initially before combining any covariates identified as being clinically important within a NMA model.

By definition, NMA uses both the within and across trial interactions. Using the across trial interaction requires the assumption of no unmeasured confounding, but unfortunately this assumption will always be hard to test. Making this assumption allows information to be gained from the network to inform both treatment effect estimates and TCI. Each trial contributes to the within trial interaction which is estimated using patient-level covariates. Meanwhile the across trial interaction is estimated through the relation of the trial-level aggregated covariates (Hong et al., 2015). Although combining within and across trial interactions can result in greater power to detect TCI, the across trial interaction can introduce ecological bias (Lambert et al., 2002). It is therefore important that the within and across trial interactions for TCI can be separated out (Hua et al., 2017). Separating out the within and across trial interactions allows the influence of the across-trial interaction on the TCI to be assessed and allows researchers to identify which data source is driving the TCI.

5. Conclusion

NMA with TCI has the potential to identify groups of patients most likely to respond to treatment. To do this NMA requires the use of both within and across trial interactions. However, the across trial interaction can be subject to ecological bias. Therefore, it is important that the within and across trial interactions in a NMA can be separated and checked for agreement. We have shown that NMA models can be parameterised to separate out the within and across trial interactions. Our proposed framework incorporates the separation of within and across trial interactions, can be applied to any outcome, outlines the steps to conducting NMA with TCI in a systematic manner, provides practical guidance for researchers and reduces the risk of unduly optimistic interpretation of treatment-covariate interactions.

Highlights

What is already known?

- Treatment-covariate interactions explore how a treatment effect varies in relation to a patient-level covariate.
- Treatment-covariate interactions contain within and across trial interactions,

where the across trial interaction is susceptible to confounding and ecological bias.

What is new?

- We propose a nine-step framework for individual participant data network meta-analysis with treatment-covariate interactions
- We use a cervical cancer example to show how to implement the framework, parameterise the network meta-analysis models to separate out the within and across trial interactions and assess the consistency of the within and across trial interactions.

Potential impact for RSM readers outside the authorsâ€™ field

- This framework provides practical guidance for researchers outlining the steps for conducting network meta-analysis with treatment-covariate interactions in a systematic manner reducing the risk of overly optimistic interpretation of treatment-covariate interactions.

Acknowledgements

The authors would like to thank the Medical Research Council for funding this research. They would also like to thank the Chemoradiotherapy for Cervical Cancer Meta-analysis Collaboration (CCCMAC) and the Neoadjuvant Chemotherapy for Cervical Cancer Meta-analysis Collaboration (NACCCMAC) who brought together the individual patient data for the meta-analyses used in the case studies, and the groups that contributed to these meta-analyses for permission to use data from their trials for this research.The contents of this publication and the methods used are, however, the sole responsibility of the authors and do not necessarily represent the views of the meta-analyses collaborative groups or the trial groups listed. CCCMAC: Gynecologic Oncology Group, USA; Yale University School of Medicine; Cross Cancer Institute and University of Alberta, Canada; Instituto de RadiologÃ-a y Centro de Lucha Contra el Cancer, Uruguay; University Medical Center Groningen and University of

Groningen, Netherlands; Institute for Oncology and Radiology of Serbia; Toronto Sunnybrook Cancer Center, Canada; First Teaching Hospital, China; AcyÂ´badem Oncology and Neurological Science Hospital, Turkey; Sanjay Gandhi Postgraduate Institute of Medical Sciences, India;Chiang Mai University, Thailand; University of Yamanashi, Japan. NACCCMAC: MRC Clinical Trials Unit, UK â€“ MRC CECA; Libera Universita "Campus Bio-Medico" di Roma, Italy; Buenos Aires University, Argentina; Royal Marsden Hospital, UK; Centro Estatal de Cancerologia, Mexico; Chang Gung Memorial Hospital, Taiwan; Istituto Nazionale per la Ricerca sul Cancro, Italy; Institut Bergonie, France; Derbyshire Royal Infirmary, UK; Tottori University School of Medicine, Japan; All India Institute of Medical Sciences, India, Hospital Pereira Rossell, Uruguay; City Hospital Birmingham, UK; Hopital General de Montreal, Canada; The Norwegian Radium Hospital, Norway; Leicester Royal Infirmary, UK; University of Sydney, Australia.

Funding

SCF received a Doctoral Training Grant from the UK Medical Research Council (MC UU 12023/21). JRC, JFT and DF are part of the London Hub for Trials Methodology Research and supported by the MRC Clinical Trials Unit Methodology Programme Grant, reference no: MC UU 12023/21. No funding body had any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Competing interests

The authors declare that they have no competing interests.

Supplementary Material

Appendix A: Additional parameterisation of within and across trial interactions

Appendix B: Sensitivity analysis - Cervical cancer NMA with treatment-stage interactions excluding patients with missing data

References

[1] Berlin, J. A., J. Santanna, C. H. Schmid, L. A. Szczech, and H. I. Feldman (2002). Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: Ecological bias rears its ugly head. *Statistics in Medicine* *21*, 371–387.

[2] Caldwell, D. M., A. E. Ades, and J. P. T. Higgins (2005). Simultaneous comparison of multiple treatments: Combining direct and indirect evidence. *BMJ* *331*, 1897–900.

[3] Chaimani, A., J. P. T. Higgins, D. Mavridis, P. Spyridonos, and G. Salanti (2013). Graphical tools for network meta-analysis in Stata. *PLOS One* *8* (10), e76654.

[4] Chaimani, A. and G. Salanti (2012). Using network meta-analysis to evaluate the existence of small-study effects in a network of interventions. *Research Synthesis Methods* *3* (2), 161–176.

[5] Chemoradiotherapy for Cervical Cancer Meta-Analysis Collaboration (2008). Reducing uncertainties about the effects of chemoradiotherapy for cervical cancer: A systematic review and meta-analysis of individual patient data from 18 randomized trials. *J Clin Oncol* *26* (35), 5802–12.

[6] Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics* *10*, 101–129.

[7] Cooper, N. J., A. J. Sutton, D. Morris, A. E. Ades, and N. J. Welton (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* (14), 1861–81.

[8] Debray, T., K. Moons, G. van Valkenhoef, O. Efthimiou, N. Hummel, R. Groenwold, and J. Reitsma (2015). Get real in individual participant data (IPD) metaâ€ • analysis: A review of the methodology. *Research Synthesis Methods* *6*, 293–309.

[9] Dias, S., A. Sutton, A. Ades, and N. Welton (2013). Evidence synthesis for decision making 2: A generalised linear modelling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making 33*, 607–617.

[10] Dias, S., N. J. Welton, D. M. Caldwell, and A. E. Ades (2010). Checking consistency in mixed treatment comparison meta-analysis. *Stat Med 29* (7-8), 932–44.

[11] Donegan, S., N. J. Welton, C. Tudur Smith, U. D'Alessandro, and S. Dias (2017). Network meta-analysis including treatment by covariate interactions: Consistency can vary across covariate values. *Research Synthesis Methods 8*, 485–495.

[12] Donegan, S., P. Williamson, U. D'Alessandro, P. Garner, and C. Tudur Smith (2013). 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* (6), 914–30.

[13] Donegan, S., P. Williamson, U. D'Alessandro, and C. Tudur Smith (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. *Statistics in Medicine 36*, 772–789.

[14] Donegan, S., P. Williamson, U. D'Alessandro, and C. Tudur Smith (2013). Assessing key assumptions of network meta-analysis: A review of methods. *Research Synthesis Methods 4*, 291–323.

[15] Efthimiou, O., T. P. A. Debray, G. van Valkenhoef, S. Trelle, K. Panayidou, K. G. M. Moons, J. B. Reitsma, A. Shang, G. Salanti, and on behalf of GetReal Methods Review Group (2016). Getreal in network meta-analysis: A review of the methodology. *Research Synthesis Methods 7*, 236–263.

[16] Fisher, D., J. Carpenter, T. Morris, S. Freeman, and J. Tierney (2017). Meta-analytical methods to identify who benefits most from treatments: Daft, deluded or deft approach?

BMJ 356, j573.

[17] Fisher, D. J., A. J. Copas, J. F. Tierney, and M. K. Parmar (2011). A critical review of methods for the assessment of patient-level interactions in individual participant data meta-analysis of randomized trials, and guidance for practitioners. *Journal of Clinical Epidemiology 64* (9), 949–967.

[18] Freeman, S. C. and J. R. Carpenter (2017). Bayesian one-step ipd network meta-analysis of time-to-event data using royston-parmar models. *Res Synth Methods, doi: 10.1002/jrsm.1253* .

[19] Govan, L., A. E. Ades, C. J. Weir, N. J. Welton, and P. Langhorne (2010). Controlling ecological bias in evidence synthesis of trials reporting on collapsed and overlapping covariate categories. *Stat Med 29* (12), 1340–56.

[20] Greenland, S. (1987). Quantitative methods in the review of epodemiological literature. *Epidemiol Rev 9*, 1–30.

[21] Günhan, B. K., T. Friede, and L. Held (2017). A design-by-treatment interaction model for network meta-analysis and meta-regression with integrated nested laplace approximations. *Res Synth Methods* .

[22] Higgins, J. P. T., D. Jackson, J. K. Barrett, G. Lu, A. E. Ades, and I. R. White (2012). Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Research Synthesis Methods 3* (2), 98–110.

[23] Higgins, J. P. T., S. G. Thompson, J. J. Deeks, and D. G. Altman (2003). Measuring inconsistency in meta-analyses. *BMJ 327* (557-560), 557.

[24] Higgins, J. P. T. and A. Whitehead (1996). Borrowing strength from external trials in a meta-analysis. *Statistics in medicine 15*, 2733–2749.

[25] Hong, H., H. Fu, K. L. Price, and B. P. Carlin (2015). Incorporation of individual-patient data in network meta-analysis for multiple continuous endpoints, with application to diabetes

treatment. *Stat Med* *34*, 2794–819.

[26] Hua, H., D. L. Burke, M. J. Crowther, J. Ensor, C. Tudur Smith, and R. D. Riley (2017). One-stage individual participant data meta-analysis models: Estimation of treatment-covariate interactions must avoid ecological bias by separating out within-trial and across-trial information. *Statistics in Medicine* *36*, 772–789.

[27] Ioannidis, J. P. (2009). Integration of evidence from multiple meta-analyses: A primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. *CMAJ* *181* (8), 488–93.

[28] Jackson, C., N. Best, and S. Richardson (2006). Improving ecological inference using individual-level data. *Statistics in Medicine* *25*, 2136–2159.

[29] Jackson, D., J. K. Barrett, S. Rice, I. R. White, and J. P. Higgins (2014). A design-by-treatment interaction model for network meta-analysis with random inconsistency effects. *Stat Med* *33* (21), 3639–54.

[30] Jackson, D., M. Law, J. K. Barrett, R. Turner, J. P. Higgins, G. Salanti, and I. R. White (2016). Extending dersimonian and laird's methodology to perform network meta-analyses with random inconsistency effects. *Stat Med* *35* (6), 819–39.

[31] Lambert, P. C., A. J. Sutton, K. R. Abrams, and D. R. Jones (2002). A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology* *55*, 86–94.

[32] Law, M., D. Jackson, R. Turner, K. Rhodes, and W. Viechtbauer (2016). Two new methods to fit models for network meta-analysis with random inconsistency effects. *BMC Med Res Methodol* *16*, 87.

[33] Lu, G. and A. E. Ades (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med* *23* (20), 3105–24.

[34] Lu, G. and A. E. Ades (2006). Assessing evidence inconsistency in mixed treatment

comparisons. *Journal of the American Statistical Association* (474), 447–459.

[35] Lu, G., N. J. Welton, J. P. T. Higgins, I. R. White, and A. E. Ades (2011). Linear inference for mixed treatment comparison meta-analysis: A two-stage approach. *Research Synthesis Methods 2* (1), 43–60.

[36] Lumley, T. (2002). Network meta-analysis for indirect treatment comparisons. *Stat Med 21* (16), 2313–24.

[37] Lunn, D., C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter (2013). *The BUGS Book. A practical introduction to Bayesian Analysis*. Texts in Statistical Science. Boca Raton, FL, USA: CRC Press.

[38] Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter (2000). Winbugs - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing 10*, 325–337.

[39] Mills, E., K. Thorlund, and J. Ioannidis (2013). Demystifying trial networks and network meta-analysis. *BMJ 346*, f2914.

[40] Morgenstern, H. (1982). Uses of ecological analysis in epidemiologic research. *Am J Public Health 72*, 1336–44.

[41] Neoadjuvant Chemotherapy for Cervical Cancer Meta-analysis Collaboration (2003). Neoadjuvant chemotherapy for locally advanced cervical cancer. *European Journal of Cancer 39* (17), 2470–2486.

[42] Piepho, H. (2014). Network-meta analysis made easy: detection of inconsistency using factorial analysis-of-variance models. *BMC Med Res Methodol 14*, 61.

[43] Riley, R. and E. W. Steyerberg (2010). Meta-analysis of a binary outcome using individual participant data and aggregate data. *Research Synthesis Methods 1*, 2–19.

[44] Riley, R. D., P. C. Lambert, J. A. Staessen, J. Wang, F. Gueyffier, L. Thijs, and F. Boutitie (2008). Meta-analysis of continuous outcomes combining individual patient data

and aggregate data. *Stat Med 27* (11), 1870–93.

[45] Rücker, G., G. Schwarzer, J. R. Carpenter, and M. Schumacher (2008). Undue reliance on I^2 in assessing heterogeneity may mislead. *BMC Med Res Methodol 8*, 79.

[46] Salanti, G. (2012). Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Research Synthesis Methods 3*, 80–97.

[47] Salanti, G., A. E. Ades, and J. P. Ioannidis (2011). Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: An overview and tutorial. *J Clin Epidemiol 64* (2), 163–71.

[48] Salanti, G., J. P. T. Higgins, A. E. Ades, and J. P. A. Ioannidis (2007). Evaluation of networks of randomized trials. *Statistical methods in medical research 17*, 279–301.

[49] Sardi, J., A. Giaroli, C. Sananes, N. G. Rueda, S. Vighi, M. Ferreira, M. Bastardas, G. Paniceres, and G. di Paola (1996). Randomized trial with neoadjuvant chemotherapy in stage IIIB squamous carcinoma cervix uteri: An unexpected therapeutic management. *Int J Gynecol Cancer 6*, 85–93.

[50] Simmonds, M. C. and J. P. T. Higgins (2007). Covariate heterogeneity in meta-analysis: Criteria for deciding between meta-regression and individual patient data. *Statistics in Medicine 26*, 2982–2999.

[51] Spiegelhalter, D. J., N. G. Best, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *J R Statist Soc B 64*, 583–639.

[52] StataCorp (2017). *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LP.

[53] Stewart, L. A., M. Clarke, M. Rovers, R. D. Riley, M. Simmonds, G. Stewart, J. F. Tierney, and PRISMA-IPD Development Group (2015). Preferred reporting items for systematic review and meta-analyses of individual participant data: The PRISMA-IPD

statement. *JAMA 313* (16), 1657–65.

[54] Thompson, S. G. and J. Higgins (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine 21*, 1559–73.

[55] Thompson, S. G. and J. Higgins (2005). Can meta-analysis help target interventions at individuals most likely to benefit? *Lancet 365*, 341–46.

[56] Tierney, J. F., C. Vale, R. Riley, C. Tudur Smith, L. Stewart, M. Clarke, and M. Rovers (2015). Individual participant data (IPD) meta-analyses of randomised controlled trials: Guidance on their use. *PLoS Medicine 12* (7), e1001855.

[57] Tu, Y. K. (2015). Using generalized linear mixed models to evaluate inconsistency within a network meta-analysis. *Value Health 18* (8), 1120–5.

[58] Tu, Y. K. (2016). Node-splitting generalized linear mixed models for evaluation of inconsistency in network meta-analysis. *Value Health 19* (8), 957–963.

[59] Tudur Smith, C., P. R. Williamson, and A. G. Marson (2005). Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. *Statistics in Medicine 24*, 1307–1319.

[60] van Valkenhoef, G., S. Dias, A. E. Ades, and N. J. Welton (2016). Automated generation of node-splitting models for assessment of inconsistency in network meta-analysis. *Res Synth Methods 7* (1), 80–93.

[61] Veroniki, A. A., H. S. Vasiliadis, J. P. Higgins, and G. Salanti (2013). Evaluation of inconsistency in networks of interventions. *Int J Epidemiol 42* (1), 332–45.

[62] White, I. R., J. K. Barrett, D. Jackson, and J. P. Higgins (2012). Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. *Res Synth Methods 3* (2), 111–25.

Table 1. Cervical cancer meta-analysis results using Royston-Parmar models. FTE = fixed treatment effect. * Values are log hazard ratios and 95% credible intervals.

Comparison	FTE*	Cochran's Q	Global non-PH test
RT v CTRT	-0.215 (-0.336, -0.086)	12.71, 15df, p=0.625	χ^2=0.161, 1df, p=0.688
RT v CT+RT short cycles	-0.191 (-0.375, -0.007)	20.69, 6df, p=0.002	χ^2=2.522, 1df, p=0.112
RT v CT+RT long cycles	0.227 (0.073, 0.385)	12.34, 10df, p=0.263	χ^2=0.006, 1df, p=0.944
RT v CT+S	-0.447 (-0.654, -0.243)	8.85, 4df, p=0.065	χ^2=0.118, 1df, p=0.731
CT+RT v CT+S	-0.444 (-0.830, -0.061)	0.01, 1df, p=0.939	χ^2=0.164, 1df, p=0.686

Table 2. Posterior mean and 95% credible intervals for treatment effects from NMA models. FTE = fixed treatment effect, RTE = random treatment effect, RT = radiotherapy, CTRT = chemoradiation, CT+RT = neoadjuvant chemotherapy plus radiotherapy, CT+S = neoadjuvant chemotherapy plus surgery. *Stage is fitted as a random trial-level effect

	FTE	**RTE**	**RTE + Stage***
CTRT	-0.211 (-0.337, -0.087)	-0.207 (-0.374, -0.046)	-0.198 (-0.346, -0.031)
CT+RT short cycles	0.028 (-0.164, 0.220)	0.086 (-0.229, 0.428)	0.005 (-0.320, 0.328)
CT+RT long cycles	0.223 (0.065, 0.380)	0.273 (0.031, 0.538)	0.254 (0.008, 0.540)
CT+S	-0.396 (-0.611, -0.185)	-0.333 (-0.701, 0.011)	-0.372 (-0.803, 0.056)
Stage			0.561 (0.475, 0.641)

Table 3. Posterior mean and 95% credible intervals for treatment and treatment-stage interaction effects from NMA models including treatment-stage interactions with within and across trial interactions separated and combined. Reference level is stages 1A-2A. RT = radiotherapy, CT+RT = neodadjuavnt chemotherapy plus radiotherapy, CT+S = neoadjuvant chemotherapy plus surgery.

	Within & across trial interactions separated	**Within & across trial interactions combined**
RT v CTRT	-0.421 (-0.910, 0.101)	-0.428 (-0.738, -0.114)
RT v CT+RT short cycles	-0.007 (-0.519, 0.550)	0.118 (-0.273, 0.596)
RT v CT+RT long cycles	0.100 (-0.551, 0.670)	0.099 (-0.426, 0.613)
RT v CT+S	0.332 (-0.593, 1.102)	-0.195 (-0.855, 0.380)
RT v CTRT - stage within	0.176 (-0.069, 0.417)	
RT v CT+RT - stage within	-0.035 (-0.285, 0.204)	
RT v CT+S - stage within	0.172 (-0.459, 0.776)	
RT v CTRT - stage across	0.165 (-0.279, 0.584)	
RT v CT+RT - stage across	0.110 (-0.315, 0.545)	
RT v CT+S - stage across	-0.563 (-1.319, 0.230)	
RT v CTRT - stage combined		0.170 (-0.043, 0.373)
RT v CT+RT - stage combined		0.006 (-0.234, 0.212)
RT v CT+S - stage combined		-0.120 (-0.635, 0.415)

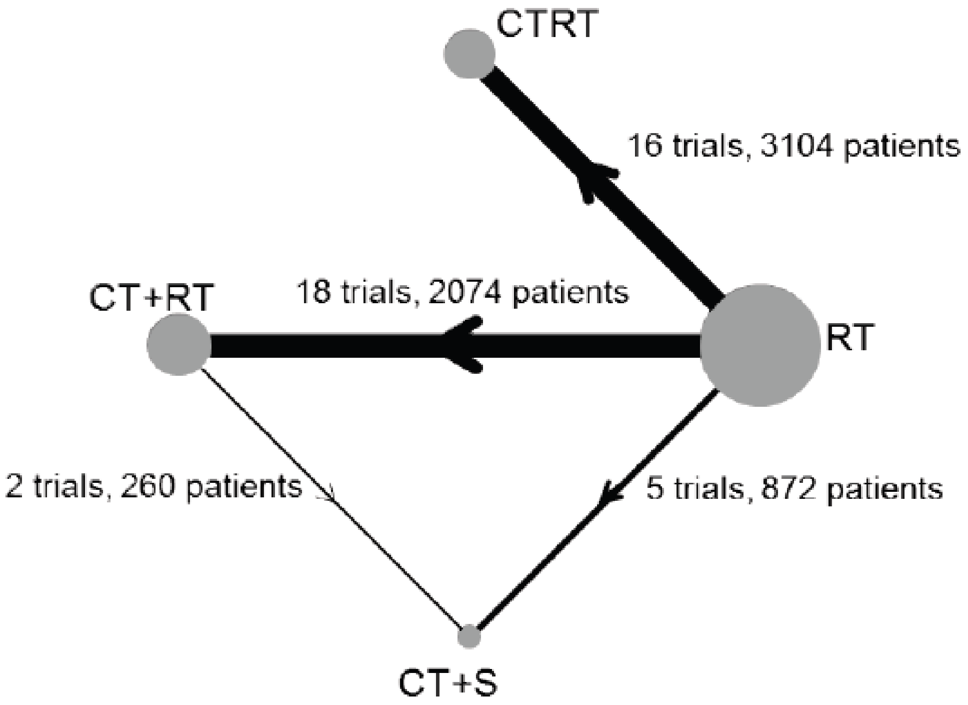

Figure 1: Cervical cancer network diagram. Node size is proportional to the number of patients randomised to each treatment and line thickness is proportional to the number of studies involved in each direct comparison. RT = radiotherapy, CTRT = chemoradiation, CT+RT = neoadjuvant chemotherapy plus radiotherapy, CT+S = neoadjuvant chemotherapy plus surgery. Note that this network diagram includes the main set of 13 RT v CTRT trials only (which in this paper is analysed as 16 trials due to the splitting of three four-arm trials each into two unconfounded comparisons of RT v CTRT), and the number of patients for each treatment arm does not add up to the total number of patients included in the network, because multi-arm patients are counted twice. There are a total of 37 trials in this network. However, in the figure the two multi-arm trials are counted three times each as they are included in the number of trials for each pairwise comparison.

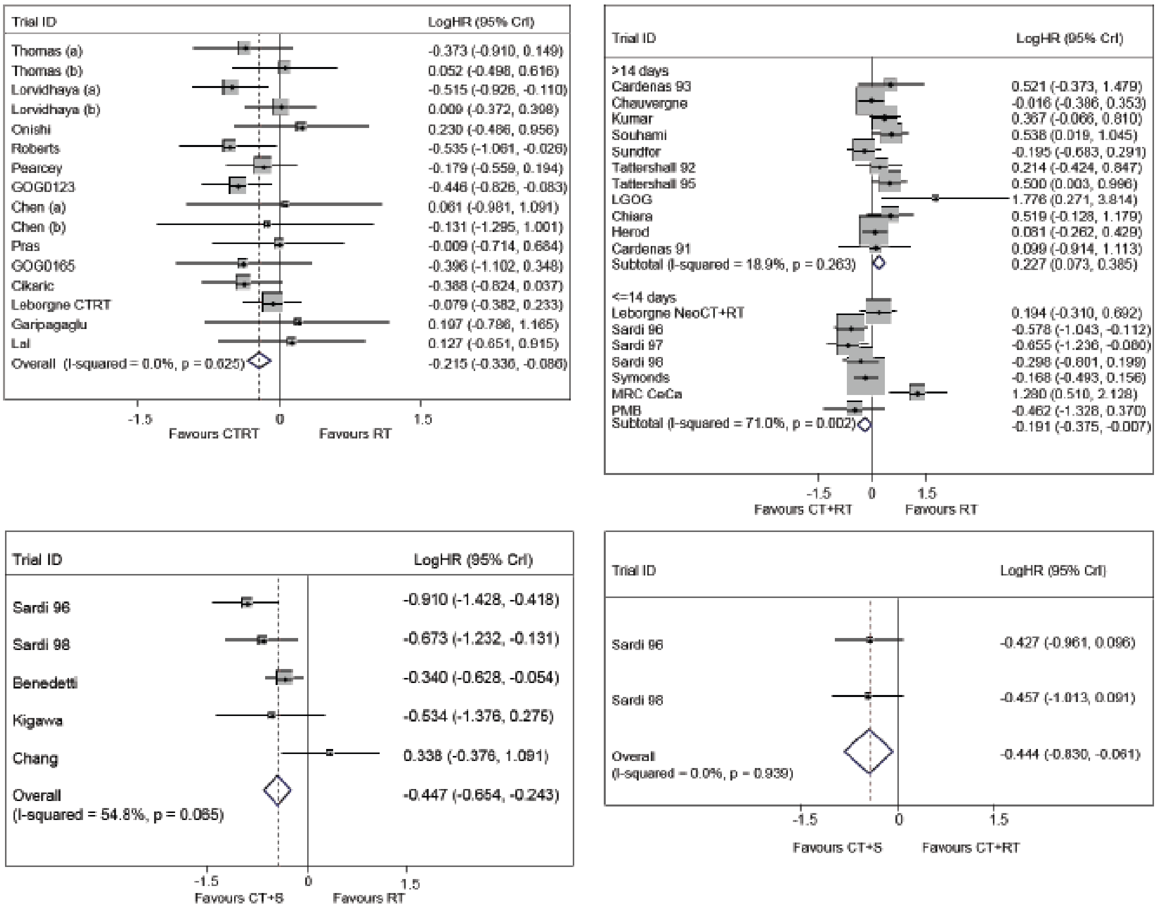

Figure 2: Cervical cancer meta-analysis plots. Trial results come from a fixed treatment effect Royston-Parmar model. Overall results come from a one-stage IPD fixed treatment effect Royston-Parmar MA model. Top left: RT v CTRT, Top right: RT v CT+RT, Bottom left: RT v CT+S, Bottom right: CT+RT v CT+S. RT = radiotherapy, CTRT = chemoradiation, CT+RT = neoadjuvant chemotherapy plus radiotherapy, CT+S = neoadjuvant chemotherapy plus surgery, LogHR = log hazard ratio, CrI = credible interval.

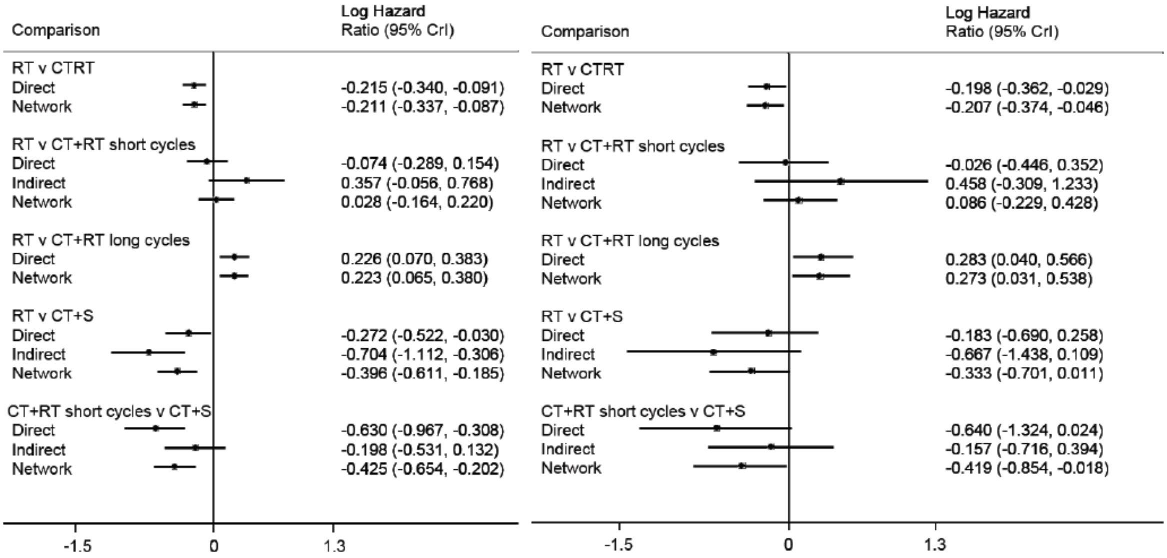

Figure 3: NMA results for the cervical cancer network. Left = Fixed treatment effect, Right = Random treatment effect. RT = radiotherapy, CTRT = chemoradiation, CT+RT = neoadjuvant chemotherapy plus radiotherapy, CT+S = neoadjuvant chemotherapy plus surgery.

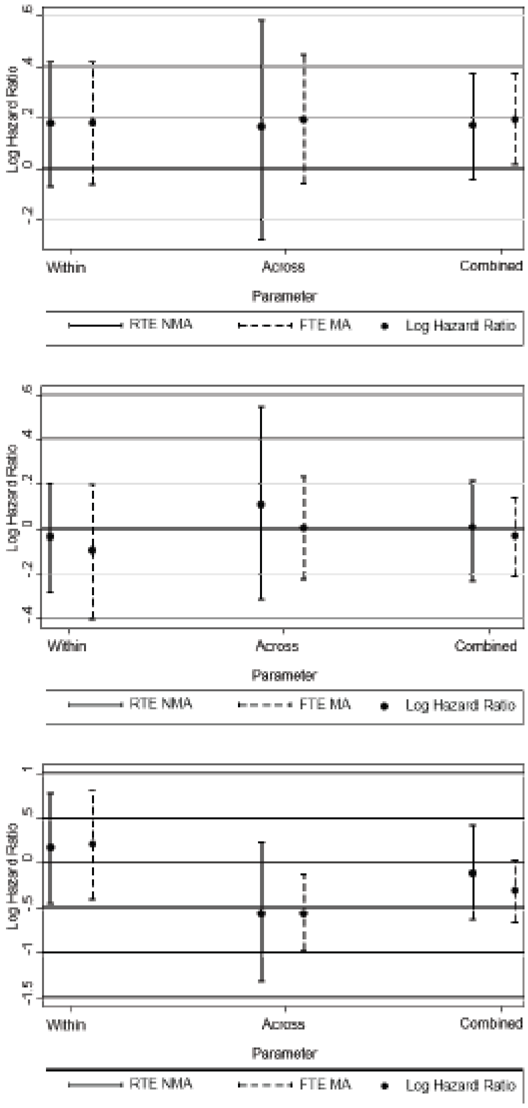

Figure 4: Treatment-stage interaction parameter estimates. Top: RT v CTRT, Middle: RT v CT+RT, Bottom: RT v CT+S. FTE = fixed treatment effect, RTE = random treatment effect, NMA = network meta-analysis, MA = meta-analysis. Solid lines represent NMA estimates. Dashed lines represent pairwise MA estimates.