
Article:
Pedder, Hugo, Boucher, Martin, Dias, Sofia orcid.org/0000-0002-2172-0221 et al. (2 more authors) (2020) Performance of Model-Based Network Meta-Analysis (MBNMA) of Time-Course Relationships:A Simulation Study. Research Synthesis Methods. pp. 678-697. ISSN 1759-2887

https://doi.org/10.1002/jrsm.1432

Received: 15 October 2019 | Revised: 28 April 2020 | Accepted: 22 June 2020

DOI: 10.1002/jrsm.1432

RESEARCH ARTICLE

Research Synthesis Methods WILEY

Performance of model-based network meta-analysis (MBNMA) of time-course relationships: A simulation study

Hugo Pedder[1] | **Martin Boucher**[2] | **Sofia Dias**[1,3] | **Margherita Bennetts**[2] | **Nicky J. Welton**[1]

[1]Department Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

[2]Pharmacometrics, Pfizer Ltd, Sandwich, UK

[3]Centre for Reviews and Dissemination, University of York, York, UK

Correspondence
Hugo Pedder, Department Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, BS8 2PS, UK.
Email: hugo.pedder@bristol.ac.uk

Funding information
Medical Research Council, Grant/Award Numbers: MR/M005232/1, MR/M005615/1; NIHR Biomedical Research Centre at University Hospitals Bristol NHS Foundation Trust; Pfizer UK; University of Bristol

Abstract

Time-course model-based network meta-analysis (MBNMA) has been proposed as a framework to combine treatment comparisons from a network of randomized controlled trials reporting outcomes at multiple time-points. This can explain heterogeneity/inconsistency that arises by pooling studies with different follow-up times and allow inclusion of studies from earlier in drug development. The aim of this study is to explore using simulation: (a) how MBNMA model parameters are affected by the quantity/location of observed time-points across studies/comparisons, (b) how reliably an appropriate MBNMA model can be identified, (c) the robustness of model estimates and predictions under different dataset characteristics. Our results indicate that model parameters for a given treatment comparison are estimated with low mean bias even when no direct evidence was available, provided there was sufficient indirect evidence to estimate the time-course. A staged model selection strategy that selects time-course function, then heterogeneity, then covariance structure, identified the true model most reliably and efficiently. Predictions and parameter estimates from selected models had low mean bias even in the presence of high heterogeneity/correlation between time-points. However, failure to properly account for heterogeneity/correlation could lead to high error in precision of the estimates. Time-course MBNMA provides a statistically robust framework for synthesizing direct and indirect evidence to estimate relative effects and predicted mean responses whilst accounting for time-course and incorporating correlation and heterogeneity. This supports the use of MBNMA in evidence synthesis, particularly when additional studies are available with follow-up times that would otherwise prohibit their inclusion by conventional meta-analysis.

KEYWORDS

MB, NMA, longitudinal, meta-analysis, NMA, simulation, time

1 | BACKGROUND

Network meta-analysis (NMA)[1,2] is commonly used to synthesize evidence from multiple studies on multiple treatments simultaneously. NMA pools evidence from all studies that form a connected network of treatment comparisons, so that inference on relative treatment effects is strengthened by combining direct head-to-head evidence with indirect evidence from the rest of the network. NMA can increase precision of estimates compared with standard pairwise meta-analysis, but it relies on the consistency assumption that there are no differences between direct and indirect treatment effects.[3,4] One reason the consistency assumption may not hold is if different studies report results at different follow-up times, and this is not accounted for in the analysis. Furthermore, the relationship between treatment efficacy and time may be of interest in itself, for example the characterization of a treatment's onset and offset of action.

A number of different methods for incorporating longitudinal data into NMA have been proposed. Riley et al[5] and Ishak et al[6] used multivariate methods to incorporate multiple follow-up times, whereas Dakin et al[7] presents a hierarchical model. Both approaches can capture differences between treatment effects at different follow-up times, but do not deliver an estimated time-course for relative treatment efficacy. To obtain an estimated time-course relationship requires a parametric model. Fractional polynomials[8] and exponential time-course functions[9] have been proposed, and more recently, model-based network meta-analysis (MBNMA), a general framework for NMA that incorporates parametric models of time-course relationships has been developed.[10]

By pooling relative effects within studies, time-course MBNMA preserves randomization and allows for testing of consistency between direct and indirect evidence in the network, whilst making use of all the available evidence at different time points. The benefit of this approach compared with NMA is that it allows inclusion of studies with a range of follow-up times, and therefore provides the possibility of including clinical trials from earlier in clinical development which may contribute valuable information on treatment efficacy. The method can be used with any parametric time-course relationship, (including exponential, E_{max}, and fractional polynomials), although because MBNMA is typically based on aggregate data only, identifiability may be an issue for models using time-course functions with two or more parameters.[10]

MBNMA was developed using a dataset of studies investigating pain relief in osteoarthritis[10] which consisted of 30 RCTs comparing 29 treatments for pain relief, measured on the Western Ontario and McMaster Universities Arthritis Index (WOMAC) scale[11] and recorded at multiple time points up to a maximum of 24 weeks (Figure S1). Following a model selection strategy, the time-course that most closely fitted the data was an E_{max} function. By modeling the time-course in this dataset using MBNMA, it was possible to include all studies and all treatments in the dataset, despite studies reporting outcomes at a range of different follow-up times. This explained significant heterogeneity and inconsistency that was present when using a single, latest follow-up time from each study, an improper approach that is sometimes used in meta-analysis.

However, this analysis raised several questions as to the statistical properties of the method. Model fit statistics were used to compare between different models, yet we were unclear of the extent to which measures such as the deviance information criterion (DIC) could be meaningfully used. MBNMA allows a range of different time-course models to be fitted, and results may be sensitive to misspecification of the underlying time-course function.

What is already known?

MBNMA is a new technique for evidence synthesis that allows incorporation of parametric time-course into NMA, which allows inclusion of studies with different follow-up times in a manner that can explain heterogeneity/inconsistency.

What is new?

This study highlights the robustness of the time-course MBNMA framework and the selection strategy that can be used to identify an appropriate model. In particular, it identifies under which conditions results from MBNMA models are likely to be of value, and in which there may be limitations.

Potential impact for RSM readers outside the authors' field

By demonstrating that time-course can be included in NMA in a statistically robust manner, we hope that this will allow the inclusion of trials from drug development into reimbursement agency decision-making. Doing so can help bridge the gap in evidence synthesis techniques that currently exist between pharmacometrics and Health Technology Appraisal.

Furthermore, results at multiple follow-up times from the same study will be correlated, with different choices for how this can be modeled. It is therefore important to assess how sensitive model results are to misspecification of the time-course function and correlation structure, and how best to select between different models.

Comparisons within the network also contained varying numbers of observations with which to inform the time-course. When estimating time-course parameters for a given non-linear function (eg, exponential, E_{max}), we expect that the number and location of observed follow-up times across studies and comparisons in the network are likely to be a critical factor in determining identifiability and the precision with which parameters can be estimated. In practice, as in the pain relief in osteoarthritis example, study follow-up times are likely to be picked to fit with a reasonable visit schedule for patients, along with the main landmark time point(s) of interest. Therefore, it is necessary to understand what impact the presence of different follow-up times will have in the estimation of the parameters in the network.

In this paper, we aim to investigate the performance of MBNMA time-course models applied to datasets generated with varying characteristics. We divide this paper into two related simulation studies which aim to answer:

1. How are MBNMA model parameters affected by the quantity and location of observed time points across studies and comparisons in the network?
2. How reliably is an appropriate time-course MBNMA model identified?
3. How robust are model estimates and predictions under different dataset characteristics?

The results from these studies can help identify in which circumstances these models can be expected to perform well, and in which they might perform poorly, allowing the robustness of conclusions drawn from time-course MBNMA to be considered in light of the number/location of observations reported in the data, the assumptions made within the modeling process, and the purpose for which the model will be used (ie, whether time-course parameters or predicted means are of interest). We begin by describing the time-course MBNMA model. We then describe the methods used for the two simulation studies, before presenting results and conclusions.

2 | TIME-COURSE MODEL-BASED NETWORK META-ANALYSIS

We briefly explain methods for time-course MBNMA. A more detailed explanation can be found in Pedder et al.[10]

2.1 | Likelihood

We assume we have a summary outcome, such as mean outcome or log-odds of response, $y_{i,\,k,\,m}$, together with standard errors, $se_{i,\,k,\,m}$, reported for each study i, arm $k = 1, ..., K_i$, and at time point $m = 1, ..., M_i$, where study i has K_i arms and reports outcomes at M_i time points. We let $s_{i,\,m}$ be the actual time at which the mth time point in study i was observed. The treatment given in study i, arm k, is indicated by $t_{i,\,k}$.

We assume the summary outcome has been transformed onto a scale where a Normal likelihood is appropriate:

$$y_{i,k,m} \sim N(\theta_{i,k,m}, se_{i,k,m})$$

in which $\theta_{i,k,m}$ is the modeled outcome (eg, predicted mean on the relevant scale) at time point m in arm k of study i.

However, when we have repeated measures from the same individuals within each study, the observations may be correlated, which can be captured with a multivariate Normal likelihood:

$$\boldsymbol{y}_{i,k} \sim \mathrm{MVN}(\boldsymbol{\theta}_{i,k}, \boldsymbol{\Sigma}_{i,k})$$

where $\boldsymbol{y}_{i,\,k}$ is a vector of the observed summary measures across the time points measured in that trial, $\boldsymbol{\theta}_{i,\,k}$ is a vector of modeled outcomes, and $\boldsymbol{\Sigma}_{i,k}$ is an $M_i \times M_i$ covariance matrix:

$$\boldsymbol{\Sigma}_{i,k} = \begin{pmatrix} se_{i,k,1}^2 & \rho_{i,k,1,2} se_{i,k,1} se_{i,k,2} & \cdots & \rho_{i,k,1,M_i} se_{i,k,1} se_{i,k,M_i} \\ \rho_{i,k,1,2} se_{i,k,1} se_{i,k,2} & se_{i,k,2}^2 & \cdots & \rho_{i,k,2,M_i} se_{i,k,2} se_{i,k,M_i} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{i,k,1,M_i} se_{i,k,1} se_{i,k,M_i} & \cdots & \cdots & se_{i,k,M_i}^2 \end{pmatrix}$$

where ρ_{i,k,m_1,m_2} is the within-study correlation between summary measures at time points m_1 and m_2 for study i arm k.

Common correlation parameters across arms and studies are typically assumed in order to improve identifiability. Furthermore, some constraints on the covariance structure are required, such as assuming a compound symmetry (CS) or autoregressive (AR1) covariance structure.[10]

2.2 | Time-course model

We put the time-course model on the aggregate-level means:

$$\theta_{i,k,m} = f(s_{i,m}, \boldsymbol{\lambda}_{i,k})$$

where f defines a functional relationship over time $s_{i,m}$, and $\boldsymbol{\lambda}_{i,\ k} = (\lambda_{0,i,k}, \lambda_{1,i,k}, \lambda_{2,i,k}, \ldots)$ are a set of parameters that describe the relationship in mean outcomes over time.[10] In all time-course models there will be a "nuisance parameter" $\lambda_{0,i,k}$ which represents the "intercept" at time $s = 0$. We put our modeling assumptions on the remaining parameters, $\lambda_{1,i,k}$, $\lambda_{2,i,k}$, ..., leaving the $\lambda_{0,i,k}$ unconstrained (achieved in a Bayesian analysis by giving independent vague prior distributions to the $\lambda_{0,i,k}$ parameters).

For example, for a two-parameter $E_{\max}$ time-course function there are two time-course parameters, and a baseline response ($\lambda_{0,i,k} = E_{0,i,k}$ - the mean response at baseline in arm k of study i):

$$f(s_{i,m}, \boldsymbol{\lambda}_{i,k}) = E_{0,i,k} + \frac{E_{\max,i,k} \times s_{i,m}}{ET_{50,i,k} + s_{i,m}} \tag{1}$$

$E_{\max,\ i,\ k}$ (equivalent to $\lambda_{1,i,k}$) is the maximum mean difference from baseline in arm k of study i and $\mathrm{ET}_{50,i,k}$ (equivalent to $\lambda_{2,i,k}$) is the time at which 50% of the maximal effect has been reached in arm k of study i.

2.3 | Network meta-analysis

The network meta-analysis (NMA) model describes the impact of treatments on one or more of the parameters of the time-course model, $\lambda_{1,i,k}$, $\lambda_{2,i,k}$, ... If the NMA model is given for a single time-model parameter, $\lambda_{1,i,k}$, we have:

$$g(\lambda_{1,i,k}) = \mu_{1,i} + \delta_{1,i,k}$$

for a given link function g which transforms the outcome to a scale where relative treatment effects may be expected to be additive. $\mu_{1,i}$ is the time-course model parameter (on the transformed scale) for arm 1 of study i, and $\delta_{1,i,k}$ the study-specific relative effect for the treatment used in arm k relative to arm 1 of study i.

For a two-parameter time-course function such as the $E_{\max}$ model we put the NMA model for the $E_{\max}$ parameter, $\lambda_{1,i,k}$, on the natural scale and the NMA model for the ET_{50} parameter, $\lambda_{2,i,k}$, on the log-scale to ensure that it can only take positive values:

$$\lambda_{1,i,k} = \mu_{1,i} + \delta_{1,i,k}$$
$$\log(\lambda_{2,i,k}) = \mu_{2,i} + \delta_{2,i,k}$$

RCTs provide comparative evidence between treatments and so our focus is on the estimation of relative effects between treatments. In these circumstances $\mu_{1,\ i}$ and $\mu_{2,\ i}$ are handled as nuisance parameters and given independent vague prior distributions in a Bayesian analysis to allow them to be unconstrained.[1,2]

Treatment effects on each time-course parameter can be either assumed "common" (often called "fixed" in meta-analysis literature) or "random" (sometimes referred to as "exchangeable") across studies. For the random effects model, study-specific treatment effects are assumed to be normally distributed around a mean treatment effect that adheres to the consistency relationships, with common between-studies variance τ^2 across treatment comparison. For a two-parameter time-course function with random effects on both time-course parameters this would be as follows:

$$\begin{aligned} \delta_{1,i,k} &\sim N\left(d_{1,t_{i,k}} - d_{1,t_{i,1}}, \tau_1^2\right) \\ \delta_{2,i,k} &\sim N\left(d_{2,t_{i,k}} - d_{2,t_{i,1}}, \tau_2^2\right) \end{aligned} \tag{2}$$

The consistency relationships reflect the comparison made between the treatment $t_{i,k}$ used on arm k and the treatment $t_{i,1}$ used on arm 1 of each study. The common effect model for each time-course parameter is obtained by setting $\tau_1^2 = 0$ or $\tau_2^2 = 0$ respectively.

The model estimates "basic parameters" $d_{1,1,k}$ and $d_{2,1,k}$, the pooled mean relative effect for treatment k relative to treatment 1 (the reference treatment for the NMA) for each time-course parameter. All other relative effects for treatment k relative to treatment c, $d_{1,c,k}$ and $d_{2,c,k}$, can then be derived from the consistency relationships[2,12]:

$$\begin{aligned} d_{1,c,k} &= d_{1,1,k} - d_{1,1,c} \\ d_{2,c,k} &= d_{2,1,k} - d_{2,1,c} \end{aligned} \tag{3}$$

2.4 | Simulation study methods

Simulation scenarios were motivated by a time-course MBNMA used to analyze a dataset of pain relief in osteoarthritis.[10] We conducted two separate simulation studies to evaluate our research questions:

1. Simulation Study I
    i. How are MBNMA model parameters affected by the quantity and location of observed time points across studies and comparisons in the network?
2. Simulation Study II
    i. How reliably is an appropriate time-course MBNMA model identified and how robust are model estimates and predictions when different model selection strategies are used, under different:
        a. Covariance structures
        b. Levels of correlation between observations
        c. Levels of heterogeneity

In Study I we explore the data requirements to fit an E_{max} MBNMA model to studies forming a closed network of three treatments by varying quantity and location of time points within studies. The results of this were then used to define scenarios with different time points in Study II with which to explore the performance of different model selection strategies on data with different covariance structures and different degrees of correlation between observations and heterogeneity (Figure 1).

Simulation protocols for each study were developed following the Aims, Data-generating mechanisms, Methods, Estimands, Performance measures (ADMEP) approach.[13] In this section we first describe the data-generating mechanisms used for all simulations, before describing aspects specific to the two simulation studies. We then describe the different models fitted in the two simulation studies, the performance measures that were computed, and the implementation.

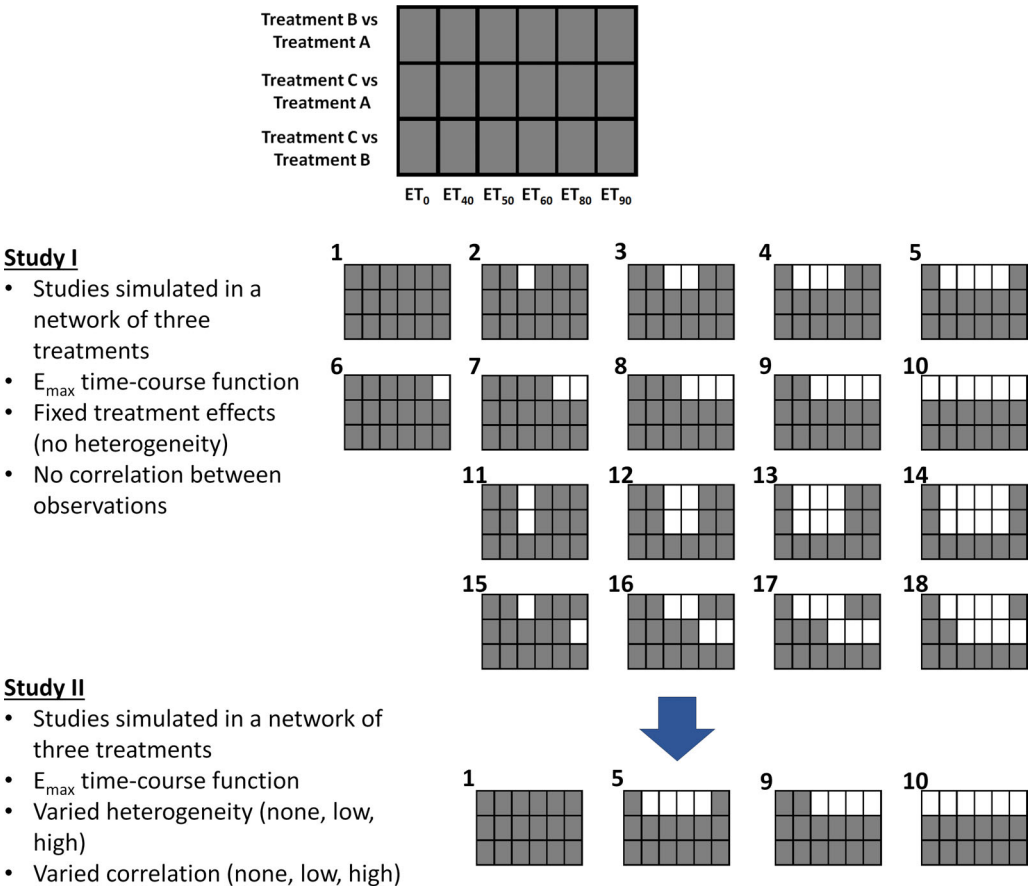

FIGURE 1 Illustrates how datasets were generated with observations present at different time points. Within each matrix, each row represents a different treatment comparison (one for each of the three in the network), and each column represents a different time point (calculated as ET, the time at which a percentage of the maximum response is achieved). Within a particular time point pattern, shaded cells represent observations that are present and white cells represent observations that are not present. If an entire row is white, this indicates studies have been removed for that treatment comparison. Time point removal patterns have been numbered to aid reference in the paper. In patterns 2 to 10, time points are removed from only a single treatment comparison, whilst in patterns 11-18 time points are simultaneously removed from two treatment comparisons. The arrow between the two studies indicates that results from Study I helped inform the design of Study II, and led to the selection of patterns 1, 5, 9 and 10 for further comparison when investigating the impact of different degrees of heterogeneity and correlation between time points [Colour figure can be viewed at wileyonlinelibrary.com]

2.5 | Dataset-generating mechanisms for all simulations

All generated datasets contained multiple two-arm studies that together formed a closed loop of three treatments, A (the network reference treatment), B and C. Four studies compared each treatment pair (A vs B, B vs C and A vs C), giving a total of 12 studies. For each study, the aggregate-level means from each arm were generated at six time-points (Supplementary Figures and Tables) based on a two-parameter $E_{\max}$ time-course function (1). The $E_{\max}$ model was used to generate all datasets as it is a flexible family of curves, commonly used for modeling time-course in pharmacometrics and clinical pharmacology, with clearly interpretable parameters. Since it contains more than one time-course parameter, it also allows investigation of the relationship between multiple time-course parameters. We specify relative treatment effects on both $ET_{50,i,k}$ and $E_{\max,i,k}$ that adhere to consistency relationships (3).

Values of each parameter used in the simulation are given in Table 1. Treatment effects for ET_{50} are given on the natural logarithmic scale to ensure absolute ET_{50} values are positive.

In preliminary work, we investigated varying the SE for study means. We found that results from datasets generated with higher SEs (lower precision) typically followed a similar pattern to results from datasets generated with lower SEs (higher precision), but with more uncertainty in estimates and higher Markov chain Monte Carlo (MCMC) convergence failure. We calculated SEs based on a coefficient of variation, $\left(\frac{SE_{i,k,m}}{\theta_{i,k,m}} \times 100\right) = 0.5\%$. This was slightly lower than the coefficient of variation found in the pain relief in osteoarthritis dataset[10] (median: 3.14%; range: 1.32%-8.55% across all observations), yet we wanted to obtain a low level of MCMC convergence failure in order to be able to evaluate performance measures.

2.5.1 | Study I

For Study I observations were generated from an $E_{\max}$ time-course function with common effects on both $E_{\max}$ and ET_{50} parameters and no residual correlation between time points. In order to investigate how the presence of observed data at different follow-up times may affect estimation of a time-course MBNMA model, observations were removed from studies in different comparisons in the following patterns to generate different datasets. These patterns were designed to illustrate cases when we have limited or no direct evidence, but some indirect evidence on different time-course parameters. We have referred to the patterns within figures pictorially using a grid system (Figure 1).

TABLE 1 Parameters and their values used in all datasets and the interpretation of those parameter values

Parameter value	Interpretation
$d_{E_{\max}A,B} = -5$	The maximum effect for treatment B is 5 points less than for treatment A
$d_{E_{\max}A,C} = -15$	The maximum effect for treatment C is 15 points less than for treatment A
$d_{ET_{50}A,B} = -0.2$	The effect of treatment B is to reduce the time at which 50% of the maximum effect is observed by $\exp(-0.2) = 0.819$ when compared with treatment A
$d_{ET_{50}A,C} = -0.5$	The effect of treatment C is to reduce the time at which 50% of the maximum effect is observed by $\exp(-0.5) = 0.607$ when compared with treatment A
$\mu_{E_{\max},i} \sim N(-40,1)$	The mean maximum response for the treatment in arm 1 of each study is normally distributed around a mean of −40 and variance of 1.
$\mu_{ET_{50},i} \sim N(\log(2.5),0.0001)$	The mean time at which 50% of the maximum response for the treatment in arm 1 of each study is observed is normally distributed around a mean of $\log(2.5) = 0.916$ and variance of 0.0001.

Note: Under the consistency assumption, $d_{E_{\max}B,C} = d_{E_{\max}A,C} - d_{E_{\max}A,B} = -10$ and $d_{ET_{50}B,C} = d_{ET_{50}A,C} - d_{ET_{50}A,B} = -0.3$, which implies that for treatment C the time at which 50% of the maximum effect is observed is reduced by $\exp(-0.3) = 0.741$ when compared with treatment B.

Based on an expected empirical SE of 0.5 for $d_{E_{\max}}$ and 0.05 for $d_{ET_{50}}$, we calculated that 5000 simulations would result in a Monte Carlo SE (MCSE) of 0.005 for $d_{E_{\max}}$ and 0.0005 for $d_{ET_{50}}$, which was more than sufficient for our investigations.

2.5.2 | Study II

In order to investigate the impact of fitting different time-course MBNMA models and using different strategies to select between them, datasets were generated using different combinations of the following characteristics to produce 15 different data-generating models:

- Different covariance structures
  - Compound symmetry (CS)
  - Autoregressive (AR1)

- Different degrees of correlation between observations. As the interpretation of correlation coefficients changes depending on the covariance structure, we selected values for ρ_{AR1} that had a mean correlation coefficient for all time points equal to ρ_{CS}.
  - High correlation ($\rho_{CS} = 0.7$, $\rho_{AR1} = 0.924$)
  - Moderate correlation ($\rho_{CS} = 0.2$, $\rho_{AR1} = 0.699$)
  - No correlation
- Between-study SD (τ)
  - Common treatment effects on E_{max} and ET_{50} ($\tau_{E_{max}} = 0$ and $\tau_{ET_{50}} = 0$)
  - Random treatment effects on E_{max} with moderate heterogeneity ($\tau_{E_{max}} = 1$) and common treatment effects on ET_{50} ($\tau_{ET_{50}} = 0$)
  - Random treatment effects on E_{max} with high heterogeneity ($\tau_{E_{max}} = 5$) and common treatment effects on ET_{50} ($\tau_{ET_{50}} = 0$)

These models were then applied to four different sets of included studies, selected based on results from Study I (Figure 1).

In total this produced 60 different datasets for Study II.

Given that there were many more datasets generated for Study II than for Study I, we examined using fewer simulations to decrease computational time. Based on an expected empirical SE of 0.5 for $d_{E_{max}}$ and 0.05 for $d_{ET_{50}}$, 742 simulations would be expected to result in a Monte Carlo SE (MCSE) of 0.013 for $d_{E_{max}}$ and 0.0013 for $d_{ET_{50}}$, which was sufficient for our investigations.

2.6 | Analysis for all simulations

The following estimands were used in both studies:

- The relative treatment effects of treatments B and C compared to the network reference treatment (A) for the different time-course parameters for E_{max} models (ET_{50} and E_{max}): $d_{E_{max}A,B}$, $d_{E_{max}A,C}$, $d_{ET_{50}A,B}$ and $d_{ET_{50}A,C}$
- The predicted mean responses at 2, 6 and 12 weeks follow up for treatments B and C ($\theta_{B,2}$, $\theta_{B,6}$, $\theta_{B,12}$, $\theta_{C,2}$, $\theta_{C,6}$, and $\theta_{C,12}$), were derived by applying the estimated relative effects to the following assumed absolute parameter values on reference treatment A[10,14]: $E_0 = 100$, $E_{max} = -40$ and $ET_{50} = \log(2.5)$. This allowed for comparison of performance measures between models with different time-course functions.

The posterior median was used as the central measure for each parameter, and the posterior SD as an indicator of precision.

2.6.1 | Study I

For Study I, the focus was on identifying how the estimation of time-course parameters was affected by the removal of different time points, given correct model specification. We therefore used the same model for analysis as was used to generate the data.

2.6.2 | Study II

For Study II, 15 different models were used for analysis (Table 2). The following model fit statistics were calculated for each analysis model, and are described in more detail in Supplementary Methodology:

- The posterior mean of the residual deviance ($\bar{D}_{res}$)
- The posterior mean of the deviance ($\bar{D}$)
  - The effective number of parameters, calculated using either the plug-in method (p_D),[15] or an approximation to the effective number of parameters (p_v).[16]
- The Deviance Information Criterion (DIC), calculated using two different approaches to compare their performance for model selection:

$$DIC_D = \bar{D}_{res} + p_D$$

$$DIC_v = \bar{D} + p_v$$

2.7 | Performance measures for all simulations

For each parameter of interest, we calculated three measures of performance. Bias was calculated to establish how reliably the posterior median targets the true parameter value. It can be expressed either as an absolute value or as a % of the true parameter value, thereby facilitating comparisons between parameters on different scales. Model SE is the mean of the posterior SDs for a parameter over all the simulations and was calculated to reflect the precision of the model. % error in model SE vs empirical SE (subsequently referred to as "% error in SE") was calculated to identify how reliably the posterior SD targets the long-run SD of the posterior median and it is therefore a measure of how reliably a model captures the "true" degree of precision in the data. Positive values reflect an underestimation of precision, whilst negative values reflect an

TABLE 2 Different models used for analysis of datasets generated for Study II

Likelihood	Time-course	$\lambda_{1,i,k}$ treatment effects	$\lambda_{2,i,k}$ treatment effects
Univariate	E_{max}	Common	Common
Univariate	E_{max}	Random	Common
Univariate	E_{max}	Common	Random
Multivariate (CS)	E_{max}	Common	Common
Multivariate (CS)	E_{max}	Random	Common
Multivariate (CS)	E_{max}	Common	Random
Multivariate (AR1)	E_{max}	Common	Common
Multivariate (AR1)	E_{max}	Random	Common
Multivariate (AR1)	E_{max}	Common	Random
Univariate	Exponential	Common	-
Univariate	Exponential	Random	-
Multivariate (CS)	Exponential	Common	-
Multivariate (CS)	Exponential	Random	-
Multivariate (AR1)	Exponential	Common	-
Multivariate (AR1)	Exponential	Random	-

Note: For E_{max} time-course, $\lambda_{1,i,k}$ and $\lambda_{2,i,k}$ correspond to E_{max} and ET_{50} parameters respectively. For exponential time-course $f(s_{i,m}, \lambda_{1,i,k}) = E_{0,i,k} + e^{\lambda_{1,i,k} \times s_{i,m}}$, such that $\lambda_{1,i,k}$ corresponds to the rate of growth/decay. CS and AR1 indicate compound symmetry and autoregressive AR1 covariance structures respectively for models with multivariate likelihoods. The following model characteristics are defined in equations: E_{max} time-course function (1), Random treatment effects (2), Common treatment effects (2) (with $\tau^2 = 0$).

overestimation of precision. For details of their calculation, see Supplementary Methodology.

MCMC convergence failure was evaluated as an additional measure of performance, as this reflects the identifiability of parameters in the different scenarios.

When presenting the performance measures from multiple datasets simultaneously, we report the median and range across the different datasets, as these are highly skewed and we aim to show the limits of the results we have found in these datasets.

2.7.1 | Study I

Performance measures were only calculated for datasets in which >90% of the simulations successfully converged. Results for datasets with <90% convergence are likely to suffer from excessive selection bias since results can only be reported for simulations that converge successfully.

2.7.2 | Study II

Performance measures were estimated for the selected model across all simulations within a particular dataset, as evaluated by different model selection strategies, using both DIC_D and DIC_v. To select a model in each simulation, DIC between different analysis models were compared, excluding those that failed to converge. The DIC for all converged models were ordered and the model with the lowest DIC was selected. However, if several models were within 3 DIC points from the model with the lowest DIC, a specific model selection strategy was used to select between these models (Table S2). We examined how results differed depending on which of three different model selection strategies was used:

1. "Best fit": Choose the model with the best fit (lowest deviance)
2. "Simplest": Choose the simplest model that is, the one with the lowest effective number of parameters (p_D or p_v depending on whether DIC_D or DIC_v respectively was being used to compare models)
3. "Staged strategy": Pedder et al[10] proposed a staged model selection process, where at each stage the simpler model is preferred over a more complex model from a subsequent stage unless the difference in DIC to the more complex model is >3. This approach involves the following stages:
   a. Fit common effect models with different time-course functions
   b. Compare random vs common treatment effects models for the selected time-course function from (a)
   c. Compare univariate vs multivariate (with different correlation structures) likelihoods for the model selected from (b)

An advantage of the "staged strategy" selection is that fewer models need to be evaluated.

Performance measures were also calculated separately for each analysis model to demonstrate the importance of model selection and the impact of failing to properly account for important modeling characteristics such as heterogeneity or correlation. However, many of these models are of limited interest since they would never be selected by any model selection strategy. Results for these are available in the appendix (Supplementary Figures - Extended) but are also commented on briefly in the manuscript.

2.8 | Implementation

Data were simulated in R version 3.5.1[17] using the 64-bit Mersenne twister algorithm for random number generation,[18] with input seeds of 15 432, 25 432, 35 432, 45 432, 55 432, 65 432, 75 432 and 85 432 (one for each of eight different nodes of the cluster computer used for analysis).

Analysis was carried out in a Bayesian framework using JAGS[19] implemented using a development version of the MBNMAtime[20] package in R (now available on CRAN). Scripts were multi-threaded to allow simulations to take place in parallel on 8 nodes of a cluster computer (Lenovo nx360 m5 compute nodes with two 14 core 2.4 GHz Intel E5-2680 v4 [Broadwell] CPUs and 128 GiB of RAM), with each of the three MCMC chains of the analysis being run in parallel on different processors of the nodes. Different numbers of iterations were used for different time-course models depending on their complexity:

- E_{max} time-course models: 50 000 burn-in iterations; 100 000 monitored iterations
- Exponential time-course models: 30 000 burn-in iterations; 30 000 monitored iterations

Alternative MCMC algorithms can also be used for Bayesian inference, and it is likely several of these would result in more rapid convergence. We use Gibbs sampling here as it is the algorithm used in JAGS,[19] the software in which the MBNMAtime[20] package has been developed. Whilst other MCMC samplers/algorithms, such as Hamiltonian Monte Carlo,[21] can be more efficient we do not expect them to result in different numbers of successfully converged simulations since the number of sampled iterations was large, ensuring that convergence failure arose due to identifiability (eg, sparse data) rather than sampling issues.

Models were considered to have "failed" to converge if any of the parameters had $\tilde{R} > 1.2$,[22] where $\tilde{R}$ is the ratio of the average variance of draws within each MCMC chain to the variance of the pooled draws across all chains. Values close to one therefore indicate good mixing of MCMC chains.

Vague normal prior distributions ($N(0, 1000)$) were given to the basic parameters $d_{ET_{50},A,k}$, $d_{E_{max},A,k}$, $d_{\lambda,\ A,\ k}$ (where k can take either B or C) and nuisance parameters $\mu_{E_{max},i}$, $\mu_{ET_{50},i}$ and $\boldsymbol{\mu_{\lambda,i}}$. For $\mu_{ET_{50},i}$ it was necessary to ensure that they only took positive values so priors for these were specified on the log-scale. Between-study SDs were given wide uniform prior distributions ($U(0, 100)$). In models with a multivariate likelihood, ρ_{CS} and ρ_{AR1} were given a uniform prior distribution ($U(-1, 1)$).

3 | RESULTS

3.1 | Study I

3.1.1 | Convergence

A very small proportion of analyzes failed to converge for the majority of datasets in Study I (<1.46% in Grids 1-4, 6-8, 10, 11-13 & 15-17). However, when both direct evidence for the AvB comparison and indirect evidence arising from AvC and BvC were limited (Grids 14 & 18), there was insufficient information to identify the models, leading to failure to converge in all simulations. When there was insufficient direct evidence for AvB to identify parameters this comparison, indirect evidence arising from AvC and BvC was able to help inform them, resulting in low convergence failure (3.28% in Grid 5 and 5.66% in Grid 9).

As convergence failure was much greater than 10% in Grids 14 & 18, performance measures have not been calculated for these datasets as this would introduce selection bias on estimation of the performance measures.

3.2 | Performance measures

3.2.1 | Bias

Mean bias (reported as a proportion of the true parameter values) on time-course parameters $d_{E_{max},A,B}$ and $d_{ET_{50},A,B}$ was higher in datasets in which there was insufficient direct evidence for AvB to independently estimate the time-course function but where indirect evidence arising from AvC and BvC was still available (Figure 2A; Grids 5 & 9).

In all other datasets in which time points were removed, % mean bias on all time-course parameters was very low (range: −1.78% to 1.79%) in Grids 1-4 & 6-8 (Figure 2A), even when time points were simultaneously

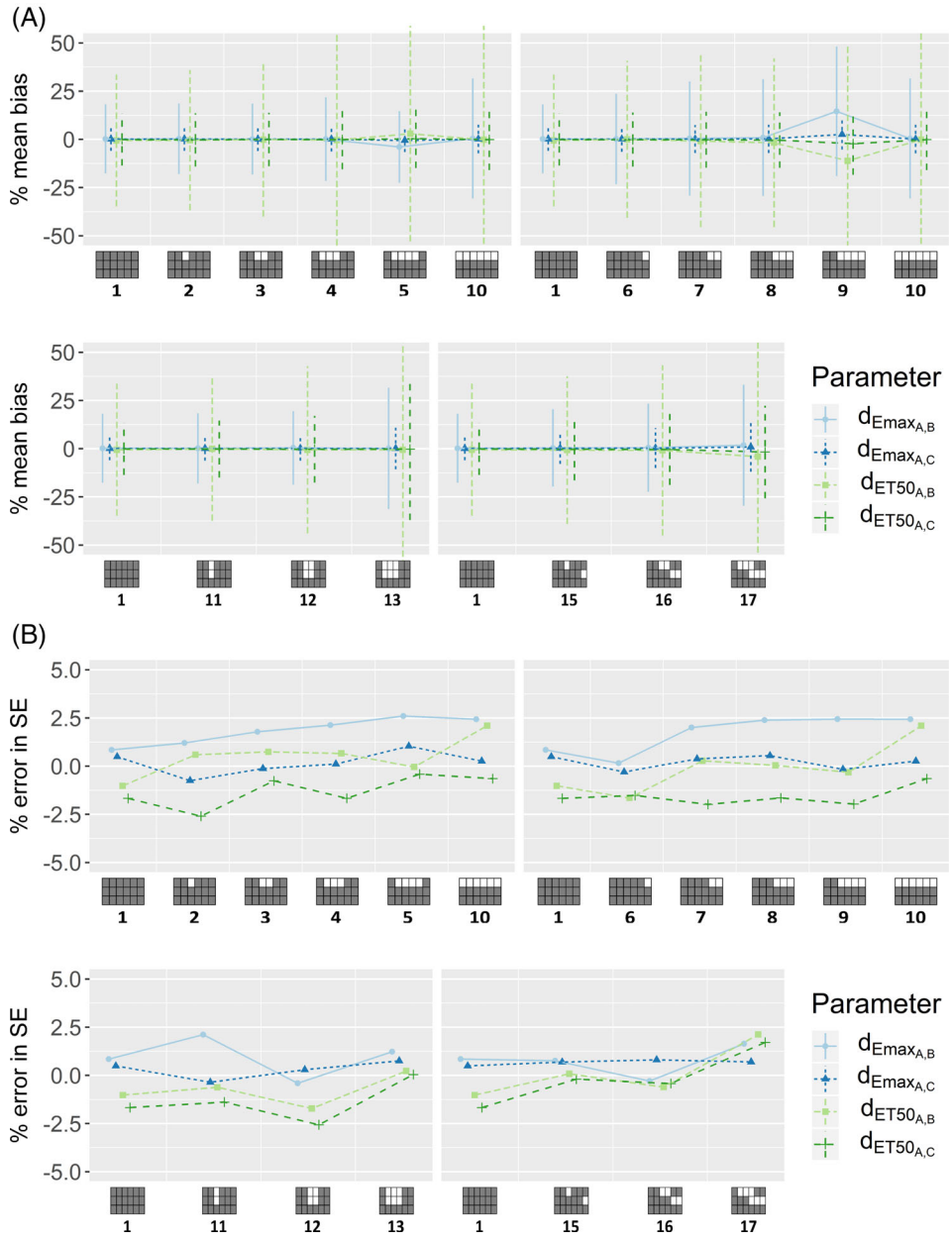

FIGURE 2 Mean bias as a % of true parameter value A, and % error in SE B, for treatment effect parameters in datasets in Study I with different patterns of study/time point removal (see Figure 1). Error bars extend symmetrically beyond y-axis limits for some points and are not visible for others where 95% CrIs are too narrow [Colour figure can be viewed at wileyonlinelibrary.com]

removed from studies in two comparisons in the network (Grids 11-13 & 15-17).

Although mean bias for predicted means followed a very similar trend to the mean bias for time-course parameters as time points were removed, mean bias as a % of the true value was much smaller, and the range of bias in simulations much lower (Supplementary Figures - Extended).

3.2.2 | Model SE vs empirical SE

Model SE increased for parameters relating to the comparison from which the time points were removed. When removing time points from studies comparing treatment B to treatment A (Grids 2-10), model SE only increased markedly for $d_{E_{\max},A,B}$ and $d_{\mathrm{ET}_{50},A,B}$ (though there was also a very slight increase for $d_{E_{\max},A,C}$ and $d_{\mathrm{ET}_{50},A,C}$) (Figures S3 and S4). However, for $d_{E_{\max},A,B}$, there was also a very clear decrease in model SE when direct information on the time-course for AvB was very limited (Grids 5 & 9), which increased again when studies comparing AvB were completely removed (Grid 10).

As with mean bias the effect on model SE of removing time points for predicted means followed a very similar trend to time-course treatment effects (Supplementary Figures - Extended).

% error in SE was generally low for all parameters in all datasets and remained stable regardless of which time points were removed from particular parts of the time-course curve in studies for any particular comparison

(range: −2.59% to 2.60%) (Figure 2B). Parameters that started with a lower error relative to other parameters continued to have a relatively lower error as time points were removed. This indicated that so long as parameters could be identified and convergence was successful, the models reliably captured the true degree of precision in the data.

3.3 | Study II

3.3.1 | Convergence

Failure to converge was low in most datasets, though failing to correctly model heterogeneity on E_{max} in datasets in which heterogeneity was present led to an increased % of simulations that failed to converge when there was limited direct evidence on one comparison (Figure S5; Grids 5 & 9). Correctly modeling this heterogeneity reduced the % of simulations that failed to converge to almost 0%. A smaller, yet opposite, effect was found in datasets in which correlation was present, where modeling correlation in the MBNMA model led to slightly higher failure to converge in Grids 5 & 9.

3.3.2 | Model selection

Where applicable, model selection methods frequently identified the different structural components of the true model from which the data were generated (Figure 3). In datasets with no residual correlation between time points, DIC_v model selection methods typically identified the true model in 91.8% (range: 7.80% to 100%) of simulations across all methods, compared with 58.3% (range: 4.44% to 100%) for DIC_D. This difference was particularly evident in datasets in which time points/studies had been removed (Supplementary Figures - Extended). However, in datasets generated with moderate heterogeneity the results were more similar, and DIC_D identified the true model in a higher % of simulations (median: 61.0%) than DIC_v (median: 47.7).

When using DIC_v as a model selection statistic, "simplest" and "staged strategy" selection methods produced very similar results. Across all datasets, they selected the same analysis model in 94.8% of simulations. These two methods often failed to select a model that accounted for heterogeneity correctly, preferring models with common treatment effects (Figure 3, in addition to figures in Supplementary Figures - Extended). For datasets generated with moderate heterogeneity, "simplest" and "staged strategy" methods with DIC_v correctly selected random treatment effects on E_{max} in 59.3% of simulations, compared to 85.6% for the "fit" method, which resulted in lower precision of estimates.

As mentioned in the Supplementary Methodology, DIC_D cannot be calculated for multivariate models which account for correlation between observations. However, DIC_v was able to select the same covariance structure as was used to generate the data in 77.1% of simulations (Figure 3, in addition to figures in Supplementary Figures - Extended), suggesting that this is a reliable statistic for comparing multivariate models in many scenarios, provided the correlation is of sufficient strength. However, in datasets generated with moderate CS covariance, DIC_v only selected a multivariate model with the correct covariance structure in 6.07% of simulations.

Selected models in all datasets and model selection methods had an E_{max} time-course function. An exponential time-course was never selected, and results of performance measures for these models are therefore not shown.

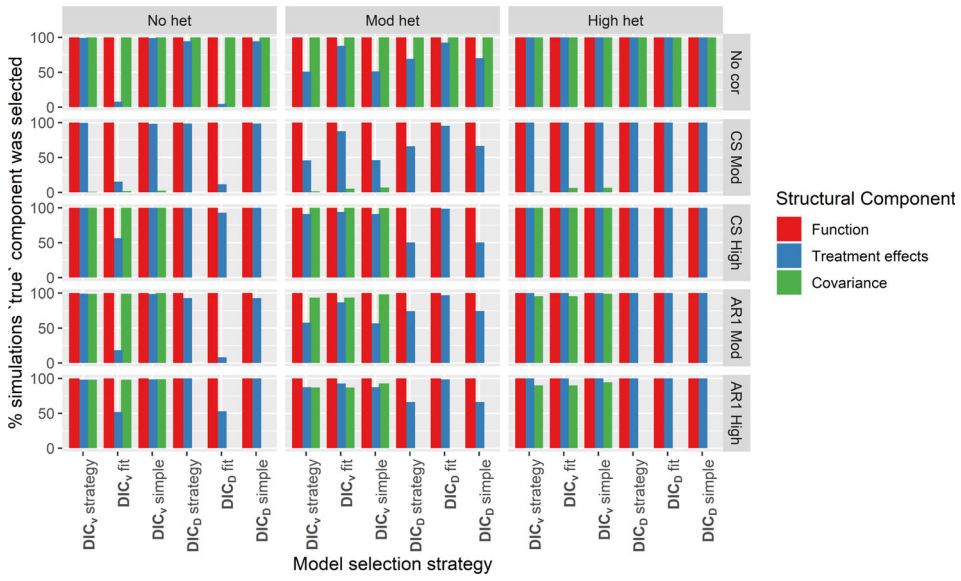

FIGURE 3 % of simulations in which different model selection strategies identified the "true" structural components (time-course function, treatment effects, covariance structure) of the model from which the data were generated with varying degrees of heterogeneity and correlation. Results are shown for DIC_D and DIC_v used as model selection statistics, with models selected based on best fit, simplest, or staged selection strategies. Results are shown for datasets in which all studies/time points are included (Grid 1) [Colour figure can be viewed at wileyonlinelibrary.com]

3.3.3 | Performance measures

We only present performance measures for the final model from each simulation here as selected by the "staged strategy" model selection method using DIC_v, since it identified an appropriate model reliably and was computationally the most efficient. Across all datasets it required fewer models to be fitted when compared with either "best fit" or "simplest" selection methods. In datasets with no heterogeneity or residual correlation between time points, as many as eight fewer models needed to be run for each simulation. Results for other model selection methods are not shown here but are available on request.

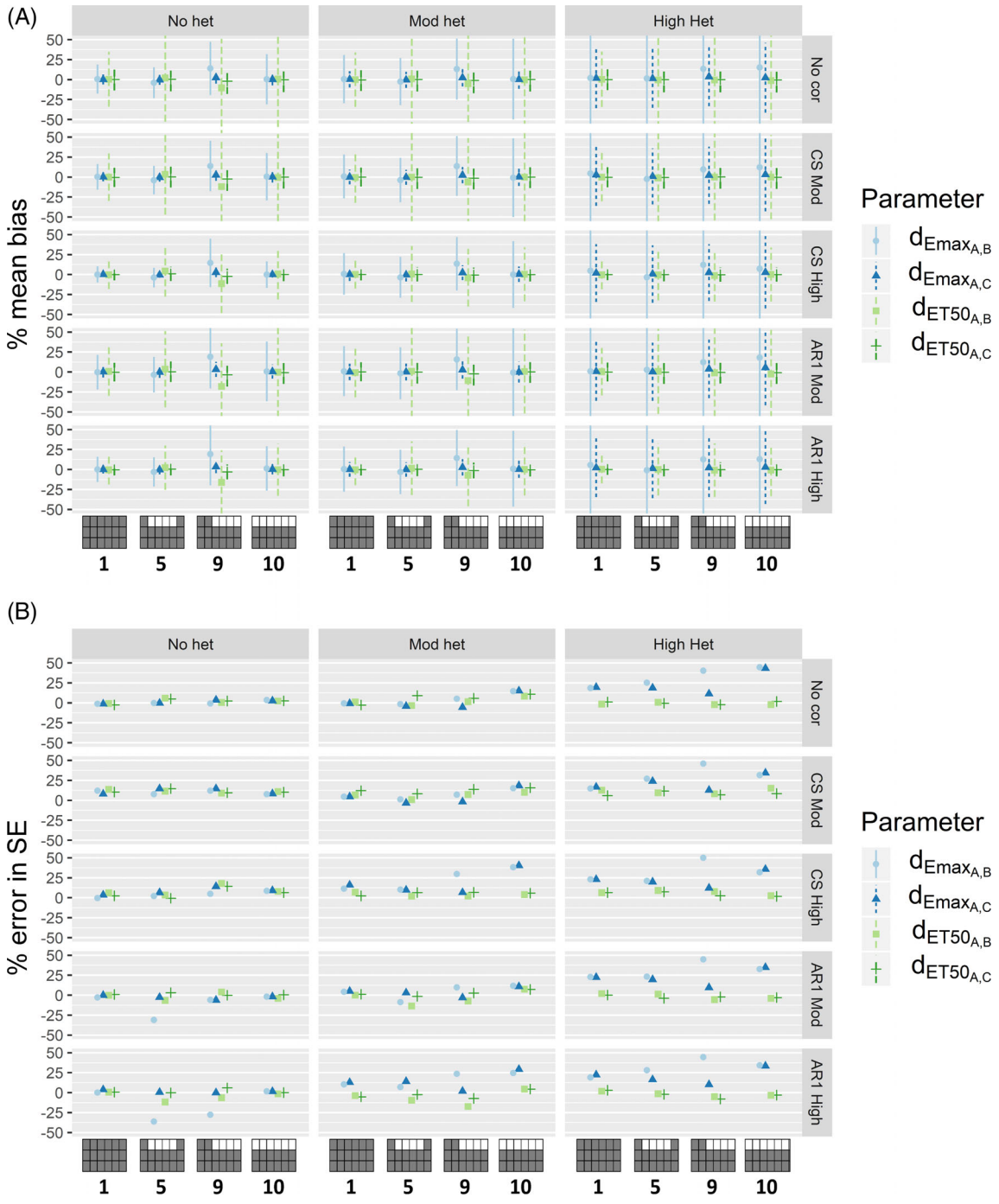

FIGURE 4 Mean bias as a % of true parameter value A, and % error in SE B, for time-course parameters from MBNMA models selected as the best using DIC_v "staged strategy" model selection in Study II. Results are presented by dataset with different heterogeneity, correlation specification and patterns of study/time point removal (see Figure 1). Error bars extend symmetrically beyond *y*-axis limits for some points and are not visible for others where 95% CrIs are too narrow [Colour figure can be viewed at wileyonlinelibrary.com]

Bias

Mean bias as a % of true parameter values on time-course parameters was low (range: −0.8% to 0.9%) in datasets with none or moderate heterogeneity and information at all time points (Figure 4A; Grid 1). This increased (range: −0.3% to 5.6%) in datasets with high heterogeneity (Grid 1). Removing studies/time points from the observed data led to increased % mean bias (Grids 5, 9 & 10). As in Study I, % mean bias was typically higher when there was limited direct evidence for AvB (Grid 9) than when these studies were removed, and time-course parameters were only informed by indirect evidence arising from AvC and BvC (Grid 10).

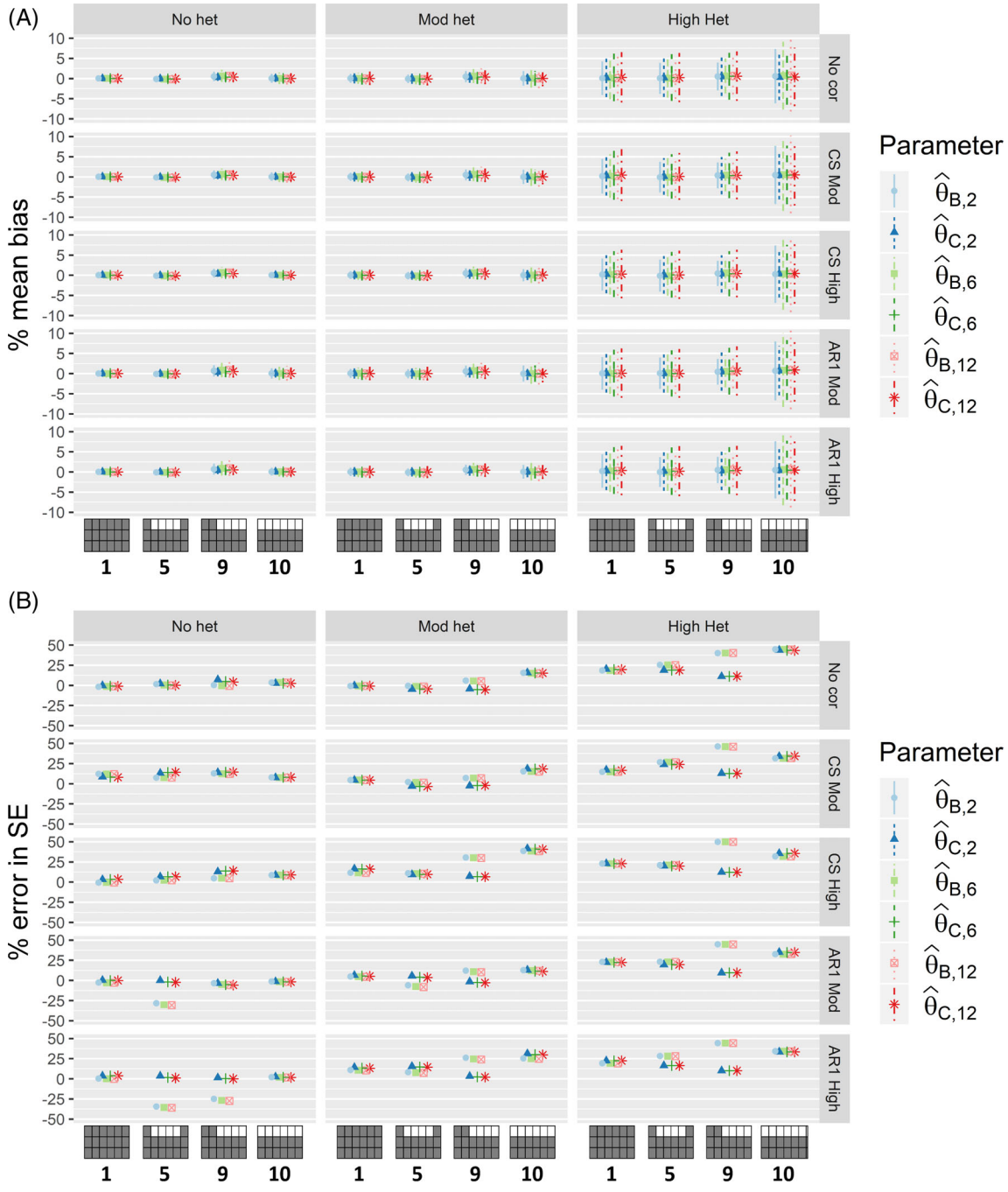

FIGURE 5 Mean bias as a % of true parameter value A, and % error in SE B, for predicted mean responses on treatments B and C, at 2, 6 and 12 weeks follow-up from MBNMA models selected as the best using DIC_v "staged strategy" model selection in Study II. Results are presented by dataset with different heterogeneity, correlation specification and patterns of study/time point removal (see Figure 1). Error bars extend symmetrically beyond y-axis limits for some points and are not visible for others where 95% CrIs are too narrow [Colour figure can be viewed at wileyonlinelibrary.com]

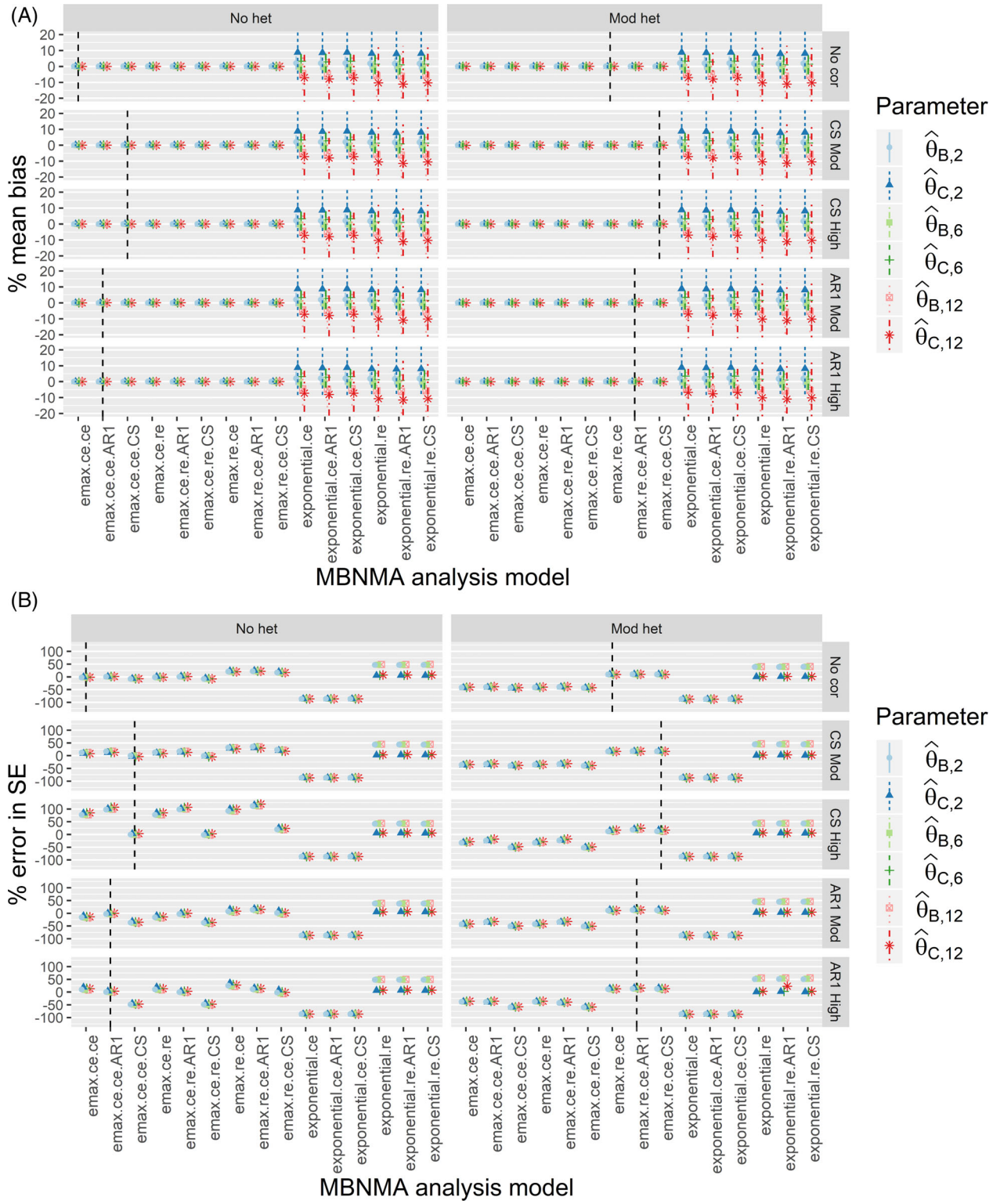

FIGURE 6 Mean bias as a % of true parameter value A, and % error in SE B, for predicted mean responses on treatments B and C, at 2, 6 and 12 weeks follow-up from different MBNMA models in datasets in which all studies/time points are present. Results are presented by dataset with different heterogeneity and correlation specification. Within the MBNMA analysis model name, the first "ce" or "re" represents common or random treatment effects respectively on E_{max} and the second represents common or random treatment effects on ET_{50}. For exponential there is only a single time-course parameter and "ce" or "re" represents common or random treatment effects on that parameter. AR1 or CS indicate that correlation has been accounted for using the respective covariance matrix structure. The true model from which the data were generated in each panel is indicated by the vertical black dashed line. High heterogeneity datasets have been excluded from results as they have high % convergence failure and so would exhibit extreme selection bias. Error bars extend symmetrically beyond y-axis limits for some points and are not visible for others where 95% CrIs are too narrow [Colour figure can be viewed at wileyonlinelibrary.com]

Results from datasets with moderate heterogeneity showed a very similar pattern of bias to those from datasets with no heterogeneity. The exception to this was in datasets with high correlation, in which case % mean bias was slightly attenuated in datasets with moderate heterogeneity compared to those with no heterogeneity in Grids 5 & 9.

The impact of heterogeneity, correlation, and the removal of studies or time points on bias followed a very similar pattern for predicted means as for time-course parameters. However, as in Study I % mean bias was substantially lower (range in % mean bias: −0.01% to 0.44% in Grid 1 datasets), and the 95%CrIs substantially narrower, implying less variability in bias (Figure 5A).

Model SE vs empirical SE

For time-course parameters and predicted means, model SE was higher in datasets with heterogeneity, and increased as time points/studies were removed from the datasets (Supplementary Figures - Extended).

% error in SE followed very similar patterns in all datasets for time-course parameters and predicted means. The results for predicted means for treatment B ($\theta_{B,2}$, $\theta_{B,6}$ and $\theta_{B,12}$) followed an almost identical pattern to $d_{E_{\max},A,B}$ and the results for predicted means for treatment C ($\theta_{C,2}$, $\theta_{C,6}$ and $\theta_{C,12}$) followed an almost identical pattern to $d_{E_{\max},A,C}$, highlighting the importance of $E_{\max}$ parameter estimation on % error in SE for making predictions.

% error in SE was always positive except in datasets generated with AR1 covariance structure when limited direct evidence on AvB was available (Grids 5 & 9). As the model SE targets the empirical SE, this suggests that these models are more likely in general to "underestimate" the precision, leading to more conservative 95% CrIs for parameters of interest. Within datasets generated with AR1 covariance structure in Grids 5 & 9 the % error in SE was more extreme for the time-course parameter for which there was more information available, and this had a corresponding effect on predicted means in the case of $E_{\max}$ parameters.

As would be expected given previous results, reduced information in the generated data (either due to removal of time points/studies for a given comparison or higher heterogeneity/correlation) led to more extreme % error in SE. However, unlike results for bias, removing studies for a comparison (Grid 10) frequently resulted in poorer performance in terms of % error in SE than when removing time points (Grids 5 and 9).

Impact of ignoring heterogeneity/correlation

Failure to properly model heterogeneity or correlation that was present in the generated data did not lead to substantial bias in either the time-course parameters or predicted means, unless an exponential time-course function was used (Figure 6A). However, there was a significant impact of ignoring either heterogeneity or correlation on % error in SE (Figure 6B).

In datasets generated with no heterogeneity, using a model for analysis with the same covariance structure as that used to generate the data led to the % error in SE being very close to zero, even when time points or studies were removed from the data. However, failing either to model the correlation, or to use the correct covariance structure, led to substantial % error in SE, particularly when the correlation was high. In datasets in which there was heterogeneity, the impact on % error in SE of failing to account correlation between observations appeared less than in datasets with no heterogeneity.

Failing to account for heterogeneity in the analysis when it was present in the generated data also had a considerable impact on % error in SE, even when the degree of true heterogeneity was only moderate. Using a common treatment effects model when a random effects model that accounted for heterogeneity on $E_{\max}$ was more appropriate led to negative error, a substantial "overestimation" of precision that led to 95% CrIs appearing tighter than they should be given the variability in the data. This effect was also exacerbated by the impact of removing studies/time points from the data (Supplementary Figures - Extended).

4 | DISCUSSION

This paper describes two studies evaluating the performance of time-course MBNMA in a series of simulated datasets of aggregate RCT responses. Study I investigated how MBNMA model parameters are affected by the quantity and location of observed time points within the dataset, whilst Study II investigated how reliably an appropriate model can be identified and how robustly the outputs can be estimated from the selected model.

4.1 | Study I

Study I illustrates that it is important to consider the quantity and location of observed follow-up times within studies in an MBNMA. We found that when there was insufficient direct evidence to be able to independently estimate the $E_{\max}$ time-course function parameters for a particular treatment comparison, it resulted in greater bias and convergence failure in the corresponding time-course parameter estimates. This was due to the difficulties of reconciling the two time-course models - a precisely estimated indirect model and a very imprecisely

estimated direct one, even though the data were simulated under the assumption of consistency between direct and indirect evidence. However, bias remained low in all scenarios for predicted means. If the objective of the synthesis is to predict the results of a potential future study or to estimate clinical results to be used in a cost-effectiveness analysis, the predicted responses are most likely to be of interest, and bias in the time-course parameters may not be of concern. On the other hand, for consultations with clinicians in which relative effects may be of greater interest, time-course parameters may be the focus of the analysis and potential bias of more concern.

The implications of our findings are that whilst we can reliably make use of indirect evidence to inform relative effects between treatments for which there is no direct evidence available, caution may have to be used if there is direct evidence available for a particular comparison that is only provided by studies that include limited time-course information.

In such a scenario, the choice of modeling approach should again depend on the objective of the analysis. If modeling the time-course is of particular importance (eg, in drug development) and there is a subset of focal treatments on which there is sufficient data, then one option may be to exclude treatments with limited data that are less crucial to decision-making. Alternatively, if estimating efficacy of all treatments simultaneously at a single time point is a priority then, provided data are available at that time point, a simple NMA should be a preferred approach as no assumption regarding the time-course relationship is required (although making this assumption can also provide additional precision even if only a single time point is of interest). However, there may be no time points at which data are available on all treatments, and we advise against "lumping" together data at different follow-up times for the purposes of synthesis, as this can introduce heterogeneity.[23]

A final option would be to model the time-course using MBNMA but to allow for sharing of information on a particular time-course parameter across treatments in the network, which may improve parameter identifiability and allow models to converge. In a time-course MBNMA of pain relief in osteoarthritis,[10] there was only direct evidence available at two observations for two treatments in the network, and for many other treatments there was limited information at earlier time points, meaning that there was insufficient evidence to inform both parameters of the E_{max} time-course model that was used. Information on the ET_{50} parameter was therefore shared across different treatments in the network to allow its estimation. Whilst this allowed for estimation of an E_{max} function, it is likely to have induced some bias in relative effects for these treatments.

The E_{max} relationship has previously been investigated in a dose-response simulation study.[24] In contrast to our results, Dutta et al[24] found that even with a wide spread of data over the E_{max} relationship, bias on E_{max} and EC_{50} (analogous to ET_{50} for dose-response relationships) was high (>15%), and it increased as the range of observations decreased. Despite this, Dutta et al[24] found similarly to our study that predicted values from the models were accurate, provided predictions were within the observed concentration range. The lower bias on E_{max} parameters found in MBNMA may be due to the added benefit of using indirect evidence and, were this evidence also to be removed from the network, convergence issues would likely be a problem before the extent of bias found by Dutta et al[24] was reached.

Understanding how the quantity of observed data and the follow-up times at which data are reported may affect estimation of time-course treatment effects also depends on which time-course parameter(s) are of interest. For an E_{max} relationship the maximum achievable response relative to competitor treatments ($d_{E_{max},c,k}$) might be considered to be the desired "target," in which case studies (contributing either direct or indirect evidence) that can provide most information to $d_{E_{max},c,k}$ will report outcomes at later follow-up times. On the other hand, for conditions in which speed of onset relative to other treatments might be more of an issue, such as migraine or illnesses in which current treatments take a long time to act (eg, psychiatric), precision and reliability in estimating $d_{ET_{50},c,k}$ may be more important. In these cases, studies that report time points closer to ET_{50} are invaluable. When considering the impact of reported follow-up times in a MBNMA, it may also be useful to consider the design of the included studies. With regards to an E_{max} time-course function, earlier phase studies are typically shorter in duration but can often include more observations. Whether E_{max} is reached in these shorter studies will depend on the onset of action of the drug and type of disease being investigated. For example, pain drugs typically have a quick onset of action and it is likely E_{max} can be well estimated in a short duration study but conversely, for drugs aimed at losing weight, E_{max} might not be well characterized in these early patient trials.

Within pharmacometrics, optimal experimental design theory seeks to identify the most important measurements required to reliably characterize a dose-response or time-course function.[25] Whilst these approaches are used when designing a study and may inform the choice of follow-up times used in the study analysis, it may be possible that the number of follow-up times collected within the study are not the same as those reported in the aggregate data, which greatly reduces their applicability in MBNMA. We urge researchers not

only to use such methods when designing a study, but also to report aggregate results at all recorded time points, as well as the correlations across these time points, to facilitate estimation of all time-course parameters in MBNMA.

4.2 | Study II

Study II makes contributions to two main areas. It firstly demonstrates the reliability of our "staged strategy" model selection method in identifying an appropriate time-course MBNMA model. This method correctly selected the model used to generate the data in a high proportion of simulations and was often able to reliably identify the time-course function, heterogeneity and covariance structure between time points. Whilst the "simplest" method was also an effective method for model selection, the benefits of the "staged strategy" method are that it considerably reduces the number of potential models that need to be fitted to identify a final model. Particularly in the case of data with no heterogeneity or residual correlation between time points, it negates the need to fit computationally intensive multivariate models and, in the case of a two-parameter time-course function, this could lead to up to eight fewer models being fitted than the "simplest" selection method without even accounting for the multitude of models that can be fitted when comparing different time-course functions. The wide range of potential MBNMA models that can typically be fitted and the computational time required to run them, particularly when using multi-parameter time-course functions and multivariate likelihoods, means that this strategy significantly facilitates the process of identifying an appropriate time-course MBNMA model.

The evaluation of different model fit statistics in Study II showed that DIC_v (calculated using p_v[16]) performs similarly to DIC_D (calculated using p_D via the plugin method[15]) for comparing random vs common treatment effect models, but has the added benefit of being calculable for multivariate likelihood models, thereby allowing comparison between univariate and multivariate models that account for correlation between time points. Whilst calculation of p_D is not possible for multivariate likelihood models due to the covariance matrix being estimated from the data, we show that p_v used in DIC_v is a reliable alternative for comparison of multivariate likelihood models with different covariance matrix structures. We therefore would recommend using DIC_v with the "staged strategy" selection method for identifying the appropriate time-course MBNMA model.

This selection method was used for MBNMA of the pain relief in osteoarthritis dataset,[10] for which there had been a question regarding whether DIC_v with the "staged strategy" was able to reliably select between univariate and multivariate likelihood models. Results from Study II confirm that this approach is likely to have selected an appropriate choice of likelihood that will have avoided substantial bias or increase in % error in SE. Even though some non-zero correlation was identified when fitting a multivariate likelihood model in this dataset, it was low, and the impact on 95% CrIs of relative treatment effects was negligible, which supported the selection of a univariate likelihood model when using the "staged strategy."

A second major contribution of Study II is that it demonstrates the robustness of model predictions and, though to a slightly lesser extent, estimation of time-course parameters. As in Study I, we found that very limited direct evidence led to greater bias on time-course parameters than in datasets in which only indirect evidence was available. Convergence was an issue in these datasets, and we suspect that this may have affected a selective sample of simulations (e.g., those with higher/lower observed values), which would therefore result in biased parameter estimates in the remaining converged simulations.

In particular, Study II highlights the importance of correctly accounting for heterogeneity and correlation between observations in MBNMA models, even when the time-course has been correctly characterized through use of an appropriate time-course function.

Previous research in meta-analysis has highlighted the importance of accounting for within-study correlations, such as when analyzing repeated measures,[26] and has also shown that ignoring correlation in model-based meta-analysis led to inflated residual variance.[27] Study II provides empirical evidence to further support this by showing that failing to account for substantial correlation present in the generated data led to increased % error in SE.

In addition, we demonstrate that the choice of covariance structure can have a considerable impact on % error in SE. Modeling using a CS covariance structure when data were generated with an AR1 structure can lead to considerable % error in SE, and vice versa. It may therefore be important to consider a variety of different covariance structure types commonly used to account for correlation between time points (eg, ARMA, Toeplitz),[28] though there may be problems with convergence for more complex covariance structures if estimation of multiple correlation coefficients is required.

Model selection methods may also struggle to select between models with different covariance structures, particularly if the true correlation is not strong. We found that in datasets with moderate CS correlation, models were typically selected that did not account for correlation between time points, and this led to slightly higher

% error in SE for predicted means and time-course parameters. It is unclear why this might be the case, why the same was not found in datasets with AR1 correlation, and how often this degree of correlation in aggregate data may occur in practice. One explanation could be due to the limitations of using DIC_v for comparison of multivariate models. An alternative approach when comparing univariate and multivariate models may be to more closely inspect the correlation parameter (if estimated from the data) to check for non-zero values.

Failing to account for heterogeneity when it was present in the generated data also led to positive % error in SE, as would be expected when modeling heterogeneous data using common treatment effects models. Yet in datasets with heterogeneity on E_{max}, even when it was modeled correctly, there was some positive % error in SE on time-course parameters and predicted means. This was most likely caused by the upwards bias in $\tau_{E_{max}}$ (Figure S6), which is a common feature when estimating heterogeneity in meta-analyses.[29,30] In this study, our choice of a conservative $U(0, 100)$ prior distribution for $\tau_{E_{max}}$ may explain the % error in SE identified in datasets with high heterogeneity.

Heterogeneity parameters are known to be sensitive to the prior distributions chosen in a Bayesian analysis, and the use of more realistic distributions such as half-normal or inverse-gamma priors may reduce bias in their estimation.[30] However, this also highlights a clear benefit of MBNMA over standard NMA in that by modeling time-course it reduces heterogeneity that might arise due to pooling of studies with different follow-up times, thereby limiting the need to estimate heterogeneity parameters and the resulting bias from doing so.

4.3 | Limitations

We have only looked at data generated from an E_{max} relationship, and one might reasonably ask whether it is fair to generalize findings from this to other time-course functions. Whilst we have not approached this in any detail (primarily due to the potentially vast number of non-linear relationships), some conclusions are likely to be more generalisable than others.

When considering the two time-course functions used for analysis in Study II, the exponential function as we defined it (Table 2) fitted the data very poorly and was never selected in any simulation. Whilst this illustrates clearly that these model selection approaches can reliably choose between different time-course functions, we would not expect to encounter this form of an exponential relationship in a pharmacometric context. We have since updated the MBNMAtime R package used for analysis in this study to incorporate a pharmacometric-specific form of the exponential function which is more generalisable to longitudinal datasets.[31]

In terms of the impact of correlation and heterogeneity, as well as the performance of model selection methods, we believe that results from this paper are generalisable to other time-course functions. However, when considering the impact of different follow-up times present in the data, the underlying time-course may lead to different conclusions.

There are also several factors that we have not considered here that are likely to be important in time-course MBNMA which could impact the external validity of the study. When analyzing longitudinal data, patient drop-out is often an important consideration. Within the modeling framework we assume either that there has been no drop-out, or if there is drop-out then either that it is missing completely at random, or that any adjustment for dropout has been accounted for already in the results reported by included studies. This approach is commonly practiced in meta-analysis due to only aggregate level data being available. An alternative approach is to restrict the inclusion criteria to studies using a specified method of imputation (eg, Last Observation Carried Forward). Whilst we have made the simplifying assumption of no drop-out in our simulations, this has allowed us to focus on the performance of the method in the ideal situation with no drop-out. Methods for investigating different missingness mechanisms have been previously described in NMA,[32,33] and these could be extended to MBNMA in future work and investigated in more detail in simulation.

We have only considered a three-treatment network, which does not provide information on the effects of "second order" indirect evidence which may add strength to improve model estimation.[34] More complex network structures/geometry (ie, the connections between treatments within the network) have been addressed previously in NMA,[35-38] and whilst this is certainly an important consideration for MBNMA, we expect that similar approaches and conclusions could be drawn.

Finally, we have not addressed the issue of inconsistency here, the potential discrepancy between direct and indirect evidence that can arise in networks of evidence.[39] Methods for identifying inconsistency in time-course MBNMA have been proposed,[10] but the potential for testing for inconsistency in network of drug development trials may be limited. In early phase studies, active treatments are all often compared to a common comparator (eg, placebo), and as trials must be internally consistent, differences between trials will manifest as heterogeneity rather than inconsistency. We hope that inconsistency in MBNMA is something that can be examined in future work using simulation, with consideration

of how model characteristics such as sharing time-course parameters across treatment comparisons may impact the ability to detect inconsistency.

5 | CONCLUSIONS

In this paper, we have demonstrated through simulation that indirect evidence can help estimate time-course parameters by providing additional information, either when limited direct evidence is available or in the absence of head-to-head trials.

We have highlighted the value of a model selection strategy for identifying an appropriate MBNMA model, and shown that DIC is a reasonable model selection statistic to use for comparison, even when it is calculated using p_v for the effective number of parameters. It also emphasizes the importance of correctly accounting for correlation between time points through the use of a multivariate likelihood with an appropriate covariance structure, and of modeling any heterogeneity present in the data.

We find that although there are some scenarios in which time-course parameter estimates may be biased, predicted responses can still be estimated reliably, which helps indicate the circumstances in which time-course MBNMA can be most useful. The true degree of precision is typically well estimated by the models provided that any heterogeneity in the data has been modeled and that correlation between time points has been appropriately accounted for.

This work demonstrates the validity of time-course MBNMA methodology and lends support to the wider application of MBNMA in evidence synthesis.

ACKNOWLEDGEMENTS

NJW, SD, and HP were supported by a grant from the MRC Methodology Research Program (ref MR/M005615/1). SD was also supported by MRC New Investigator Research Grant (MR/M005232/1). MB, MB, and HP were supported by Pfizer Ltd. NJW was partly supported by the NIHR Biomedical Research Centre at University Hospitals Bristol NHS Foundation Trust and the University of Bristol. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care.

CONFLICT OF INTEREST

The author reported no conflict of interest.

DATA AVAILABILITY STATEMENT

The simulated data that support the findings of these studies are openly available in figshare at http://doi.org/10.6084/m9.figshare.9387800.

ORCID

Hugo Pedder https://orcid.org/0000-0002-7813-3749
Sofia Dias https://orcid.org/0000-0002-2172-0221
Margherita Bennetts https://orcid.org/0000-0001-5198-5274
Nicky J. Welton https://orcid.org/0000-0003-2198-3205

REFERENCES

1. Dias S, Ades AE, Welton NJ, Jansen JP, Sutton AJ. *Network Meta-Analysis for Decision Making*. Hoboken: Wiley; 2018.
2. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med*. 2004;23(20):3105-3124.
3. Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomized controlled trials. *Med Decis Making*. 2013;33(5):641-656.
4. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. *Stat Med*. 2010;29(7–8):932-944.
5. Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Stat Med*. 2007;26(1):78-97.
6. Ishak KJ, Platt RW, Joseph L, Hanley JA, Caro JJ. Meta-analysis of longitudinal studies. *Clin Trials*. 2007;4(5):525-539.
7. Dakin HA, Welton NJ, Ades AE, Collins S, Orme M, Kelly S. Mixed treatment comparison of repeated measurements of a continuous endpoint: an example using topical treatments for primary open-angle glaucoma and ocular hypertension. *Stat Med*. 2011;30(20):2511-2535.
8. Jansen JP, Vieira MC, Cope S. Network meta-analysis of longitudinal data using fractional polynomials. *Stat Med*. 2015;34(15):2294-2311.
9. Ding Y, Fu H. Bayesian indirect and mixed treatment comparisons across longitudinal time points. *Stat Med*. 2013;32(15):2613-2628.
10. Pedder H, Dias S, Bennetts M, Boucher M, Welton NJ. Modelling time-course relationships with multiple treatments: model-based network meta-analysis for continuous summary outcomes. *Res Synth Methods*. 2019;10(2):267-286.
11. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *J Rheumatol*. 1988;15(12):1833-1840.
12. Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Med Decis Making*. 2013;33(5):607-617.
13. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. arXiv:171203198v1. 2017.
14. Dias S, Welton NJ, Sutton AJ, Ades AE. Evidence synthesis for decision making 5: the baseline natural history model. *Med Decis Making*. 2013;33(5):657-670.
15. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. *J R Statistic Soc B*. 2002;64(4):583-639.

16. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. 2nd ed. Abingdon, Oxforshire: Taylor & Francis Inc; 2003.
17. R: A language and environment for statistical computing [computer program]. Vienna, Austria, 2018.
18. Matsumoto M, Nishimura T. Mersenne twister: a 623-dimensionally equidistributed uniform psuedorandom number generator. *ACM Trans Model Comput Simul*. 1998;8(1):3-30.
19. JAGS [computer program]. Version 4.3.0, 2017.
20. MBNMAtime [computer program]. Version 0.1.0. figshare, 2019.
21. Betancourt MJ, Girolami M. Hamiltonian Monte Carlo for Hierarchical Models. arXiv 13120906. 2013.
22. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. *J Comput Graph Stat*. 1998;7(4):434-455.
23. Dias S, Sutton AJ, Welton NJ, Ades AE. Evidence synthesis for decision making 3: heterogeneity–subgroups, meta-regression, bias, and bias-adjustment. *Med Decis Making*. 2013;33(5):618-640.
24. Dutta S, Matsumoto Y, Ebling WF. Is it possible to estimate the parameters of the sigmoid E_{max} model with truncated data typical of clinical studies? *J Pharm Sci*. 1996;85(2):232-239.
25. Aarons L, Ogungbenro K. Optimal design of pharmacokinetic studies. *Basic Clin Pharmacol Toxicol*. 2010;106(3):250-255.
26. Riley RD. Multivariate meta-analysis: the effect of ignoring within-study correlation. *J R Statistic Soc A*. 2009;172:789-811.
27. Ahn JE, French JL. Longitudinal aggregate data model-based meta-analysis with NONMEM: approaches to handling within treatment arm correlation. *J Pharmacokinet Pharmacodyn*. 2010;37(2):179-201.
28. Chan JSK, Choy STB. Analysis of covariance structures in time series. *J Data Sci*. 2007;6:573-589.
29. Huedo-Medina TB, Sanchez-Meca J, Marin-Martinez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychol Methods*. 2006;11:193-206.
30. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Stat Med*. 2005;24(15):2401-2428.
31. MBNMAtime: Run Time-Course MBNMA Models. R package version 0.1.3. [computer program]. CRAN; 2020.
32. Mavridis D, Chaimani A, Efthimiou O, Leucht S, Salanti G. Addressing missing outcome data in meta-analysis. *Evid Based Ment Health*. 2014;17(3):85-89.
33. Mavridis D, White IR, Higgins JP, Cipriani A, Salanti G. Allowing for uncertainty due to missing continuous outcome data in pairwise and network meta-analysis. *Stat Med*. 2015;34(5):721-741.
34. Caldwell DM, Dias S, Welton NJ. Extending treatment networks in health technology assessment: how far should we go? *Value Health*. 2015;18(5):673-681.
35. Salanti G, Kavvoura F, Ioannidis JP. Exploring the geometry of treatment networks. *Ann Intern Med*. 2008;148:544-553.
36. Konig J, Krahn U, Binder H. Visualizing the flow of evidence in network meta-analysis and characterizing mixed treatment comparisons. *Stat Med*. 2013;32(30):5414-5429.
37. Rucker G. Network meta-analysis, electrical networks and graph theory. *Res Synth Methods*. 2012;3:312-324.
38. Salanti G, Higgins JP, Ades A, Ioannidis JP. Evaluation of networks of randomized trials. *Stat Methods Med Res*. 2008;17:279-301.
39. Lu G, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. *J Am Statist Assoc*. 2006;101:447-459.

SUPPORTING INFORMATION

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

How to cite this article: Pedder H, Boucher M, Dias S, Bennetts M, Welton NJ. Performance of model-based network meta-analysis (MBNMA) of time-course relationships: A simulation study. *Res Syn Meth*. 2020;1–20. https://doi.org/10.1002/jrsm.1432