Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2018, Vol. 27(5) 1351–1364

DOI: 10.1177/0962280216660741
journals.sagepub.com/home/smm
SAGE

An overview of methods for network meta-analysis using individual participant data: when do benefits arise?

Thomas PA Debray,[1,2] Ewoud Schuit,[1,2,3] Orestis Efthimiou,[4,5] Johannes B Reitsma,[1,2] John PA Ioannidis,[3] Georgia Salanti,[4,5,6] and Karel GM Moons[1,2] on behalf of GetReal Workpackage[4]

Abstract

Network meta-analysis (NMA) is a common approach to summarizing relative treatment effects from randomized trials with different treatment comparisons. Most NMAs are based on published aggregate data (AD) and have limited possibilities for investigating the extent of network consistency and between-study heterogeneity. Given that individual participant data (IPD) are considered the gold standard in evidence synthesis, we explored statistical methods for IPD-NMA and investigated their potential advantages and limitations, compared with AD-NMA. We discuss several one-stage random-effects NMA models that account for within-trial imbalances, treatment effect modifiers, missing response data and longitudinal responses. We illustrate all models in a case study of 18 antidepressant trials with a continuous endpoint (the Hamilton Depression Score). All trials suffered from drop-out; missingness of longitudinal responses ranged from 21 to 41% after 6 weeks follow-up. Our results indicate that NMA based on IPD may lead to increased precision of estimated treatment effects. Furthermore, it can help to improve network consistency and explain between-study heterogeneity by adjusting for participant-level effect modifiers and adopting more advanced models for dealing with missing response data. We conclude that implementation of IPD-NMA should be considered when trials are affected by substantial drop-out rate, and when treatment effects are potentially influenced by participant-level covariates.

Keywords

Meta-analysis, network meta-analysis, individual participant data, missing data, repeated measurements, mixed treatment comparison

1 Introduction

Network meta-analysis (NMA) is a common approach to synthesizing the efficacy of multiple therapeutic interventions by borrowing information from so-called indirect comparisons.[1] Because these comparisons are the result of observational inference making, their inclusion may lead to confounding and thereby generate heterogeneity and inconsistency in an NMA. This is particularly problematic when effect modification exists and the distribution of effect modifiers differs across studies.[2] Several reviews have found that about one-eighth of NMAs suffer from inconsistency,[3–5] and that lower heterogeneity is associated with higher rates of detected inconsistency.[3] Although the use of NMA meta-regression has been proposed to investigate sources of heterogeneity in treatment effects,[6,7] it is well known that this approach has limited power and is substantially prone to ecological bias when study averages are used to represent covariates that vary at the level of the individual

[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, The Netherlands
[2]Cochrane Netherlands, University Medical Center Utrecht, The Netherlands
[3]Meta-Research Innovation Center at Stanford, Stanford University, USA
[4]Institute of Social and Preventive Medicine, University of Bern, Switzerland
[5]Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Greece
[6]Institute of Primary Health Care, University of Bern, Switzerland

Corresponding author:
Thomas PA Debray, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands.
Email: T.Debray@umcutrecht.nl

Table 1. Availability of HAMD responses in the case study.

Trial	Comparison	N	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$	$l=6$
1	TeCA: Plac	90	84	78	70	67	55	47
2	TeCA: Plac	90	85	77	70	0	60	53
3	TeCA: Plac	149	144	138	106	92	80	63
4	TCA_1: Plac	84	74	67	63	60	53	51
5	TCA_1: Plac	100	93	83	70	68	54	53
6	TCA_1: Plac	100	99	95	91	87	82	78
7	TCA_1: Plac	100	95	93	89	86	66	63
8	TCA_1: Plac	280	261	251	229	223	0	194
9	TeCA:TCA_1	248	246	234	230	228	218	210
10	TeCA: Plac	113	0	0	0	0	0	83
11	TeCA: Plac	50	0	49	48	0	42	0
12	TCA_1: Plac	22	0	22	0	16	0	16
13	TeCA:TCA_2	174	0	173	0	146	0	121
14	TeCA:TCA_3	163	0	153	0	142	0	132
15	TeCA: Plac	132	0	123	0	105	97	0
16	TeCA:TCA_1	156	0	142	0	127	118	0
17	TeCA:TCA_1	205	0	0	0	0	0	143
18	TeCA:TCA_1	200	187	178	0	168	0	169

In the primary analyses, we only included trials for which HAMD responses after a follow-up of 6 weeks were available (i.e. Trials 11, 15 and 16 were excluded from the analysis).
l: follow-up time in weeks; Plac: placebo; TCA: tricyclic antidepressants; TeCA: tetracyclic antidepressants.

participants in the trials.[8–10] As a result, it is often difficult to identify sources of network inconsistency and heterogeneity in NMA that are solely based on aggregate data (AD).

It is widely accepted that the use of individual participant data (IPD) is desirable during evidence synthesis, as it may help to improve the quantity and quality of the data, to standardize outcomes across included trials and enable detailed data checking.[11,12] The use of IPD also offers more flexibility in the investigation of effect modifiers, and therefore appears to be particularly useful in addressing the inconsistencies of an NMA.[8,13] It may come as no surprise that meta-analysis of IPD has been put forward as the 'gold standard' of systematic reviews,[14] and is being increasingly implemented by medical researchers.[11,12]

Over the past few years, several methods have been proposed to perform NMA[1] and to combine multiple IPD sets.[12] Although these methods can be combined, to perform IPD-NMA, there is currently little guidance on deciding whether such complicated analyses should be initiated and how they should be conducted. For this reason, we set out to explore common challenges and potential advantages of IPD-NMA. Hereto, we present and illustrate a generic NMA framework to (a) combine IPD, (b) include covariates (prognostic factors or effect modifiers), (c) address missing response data and (d) account for longitudinal responses. With this paper, we aim to illustrate under what circumstances it might be desirable to obtain IPD, and how statistical models should be designed to integrate established best practices. While the extension will be apparent,[8,13] we do not develop methods for combining IPD and AD, as our motivating data include IPD only. Instead, we compare the IPD-NMA models with AD-NMA models that make use of commonly reported estimates of relative treatment effect.

2 Case study data

We obtained IPD from 18 randomized trials, including a total of 2456 participants, investigating the comparative (relative) efficacy of several (three tricyclic (TCA) and one tetracyclic (TeCA)) antidepressants (Table 1 and the online Appendix 1). All included randomized trials were carried out within Phase IIb, III and IV trials, and were two-armed, comparing antidepressants versus placebo or versus an active comparator (Figure 1(a)). We anonymized all trial information that could be used to identify actual drug names or manufacturers. Because the effects of the three tricyclic antidepressants were similar we considered them in a single node in the evidence network (Figure 1(b)).

All participants in all trials were diagnosed with major depressive disorder by a psychiatrist according to DSM-III (*Diagnostic and Statistical Manual of Mental Disorders*, version III) criteria and met the following

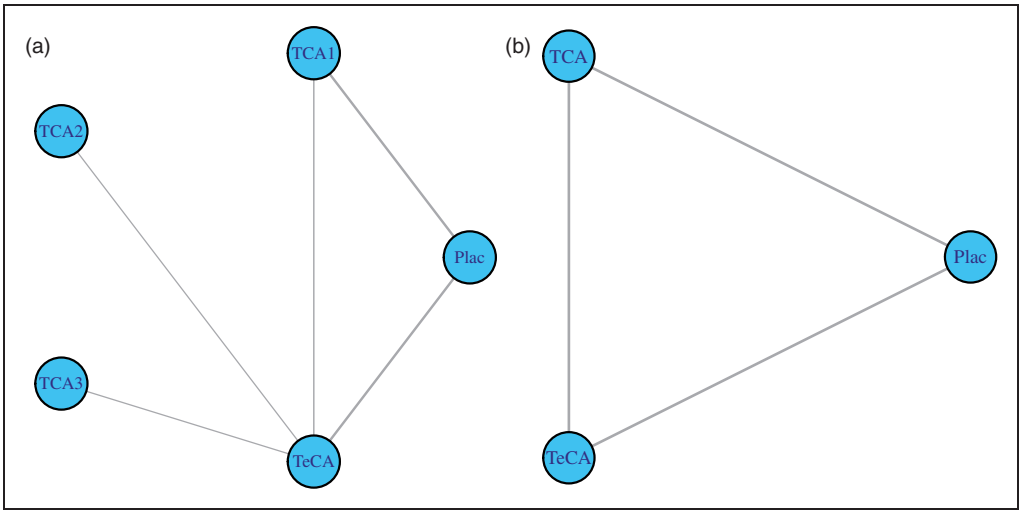

Figure 1. Network of evidence in the case study. (a) Original network of evidence. (b) Simplified network of evidence: treatments TCA1, TCA2 and TCA3 are merged into a single treatment class TCA.
Plac: placebo; TCA: tricyclic antidepressant; TeCA: tetracyclic antidepressant.

inclusion criteria: ≥18 years old, at least moderate disease severity according to the Hamilton depression (HAMD) rating scale (HAMD ≥ 16), non-smoking, non-suicidal and at least one post-baseline assessment. In general, concomitant use of psychotropic medication or psychotherapy was an exclusion criterion. For all participants, screening measurements of HAMD scores were performed. These measurements were different from the (separate) baseline measurements.

The primary response in each trial was the level of depression measured on the HAMD rating scale on a weekly basis (up to a total follow-up time of 6 weeks). The HAMD rating scale consists of 17 items, each with a possible score that ranges from 0 to 2 or 0 to 4, with larger values corresponding to more severe depression. The maximum score that can be achieved is 54. In this study, we evaluated the comparative treatment effects of TCA and TeCA antidepressants after a follow-up period of 6 weeks. In this regard, it important to note that all trials suffered from drop-out over time; missingness of responses after 6 weeks follow-up ranged from 21 to 41%. Drop-out occurred more frequently in trials that included placebo treatment as comparator (online Appendix 1).

The IPD contains information for each subject on the HAMD score (measured at $l = 1,\ldots,6$ weeks after treatment allocation) and the baseline response ($l = 0$). Let M indicate the number of studies and N_{ij} indicate the number of participants in study i receiving treatment j. We denote HAMD responses as H_{ijkl}, where i indicates the study, j the treatment, k the participant and l the follow-up time. The baseline HAMD measurement is then denoted as H_{ijk0}, and considered as a potential effect modifier. This covariate is centred by subtracting corresponding values with their overall mean, yielding x_{ijk}.

For each trial we also consider the corresponding AD. These AD may, for instance, be reported in the literature or can be generated from IPD at hand by analysing each trial separately. Let $\hat{y}_{i6}$ and $\hat{s}^2_{i6}$ denote the estimated treatment effect and corresponding error variance in terms of the mean HAMD response at week 6 in trial i. We here consider three scenarios for obtaining $\hat{y}_{i6}$ and $\hat{s}^2_{i6}$. For all scenarios, we estimate the mean difference in HAMD score between the active and baseline comparator after a follow-up period of 6 weeks using traditional regression analysis in each trial (online Appendix 2). This results in treatment effect estimates for 15 (out of 18) trials, as some trials did not report 6-week follow-up data. The results shown in Table 2 indicate that these scenarios yield roughly similar and rather imprecise estimates of treatment effect after $l = 6$ weeks.

2.1 Scenario 1

In the first scenario, we simply ignore drop-out of participants and limit the analyses to those participants with complete follow-up data for $l = 6$ weeks. This approach is still very common in the literature. For instance, a recent review indicated that about 45% of published randomized controlled trials adopt complete case analysis in the primary analysis.[15]

Table 2. Overview of generated aggregate data (AD).

Trial	Comparison	Scenario 1		Scenario 2		Scenario 3	
i	t_i:b_i	H_0	$d_{t_ib_i}$	H_0	$d_{t_ib_i}$	H_0	$d_{t_ib_i}$
1	TeCA: Plac	22.2	−3.54 (2.63)	22.1	−6.72 (1.80)	22.1	−6.07 (2.35)
2	TeCA: Plac	23.7	−1.94 (2.16)	23.6	−1.46 (1.78)	23.6	−2.22 (2.10)
3	TeCA: Plac	22.9	−0.77 (1.81)	23.1	1.59 (1.40)	23.1	1.11 (1.71)
4	TCA: Plac	24.7	−6.92 (2.11)	24.8	−5.36 (1.77)	24.8	−7.00 (2.15)
5	TCA: Plac	21.5	1.80 (1.06)	21.6	−3.00 (1.40)	21.6	−0.23 (1.25)
6	TCA: Plac	27.6	−4.67 (1.60)	27.2	−3.77 (1.57)	27.2	−4.49 (1.72)
7	TCA: Plac	23.3	−3.33 (1.79)	23.4	−5.34 (1.49)	23.4	−5.56 (1.90)
8	TeCA: Plac	21.9	−3.03 (1.06)	22.2	−2.22 (1.00)	22.2	−2.78 (1.09)
9	TeCA:TCA	25.8	0.13 (1.11)	25.9	0.02 (1.23)	25.9	0.21 (1.24)
10	TeCA: Plac	24.0	−3.46 (2.02)	24.0	−1.83 (1.73)	24.0	−3.46 (2.09)
12	TCA: Plac	29.1	−0.95 (2.44)	30.2	2.89 (4.08)	30.2	−0.50 (3.37)
13	TeCA:TCA	26.0	2.00 (1.17)	25.7	0.99 (1.34)	25.7	1.81 (1.21)
14	TeCA:TCA	22.4	0.14 (1.19)	22.3	−0.50 (1.24)	22.3	0.42 (1.16)
17	TeCA:TCA	27.2	0.82 (1.23)	26.8	1.96 (1.36)	26.8	0.82 (1.24)
18	TeCA:TCA	24.7	0.42 (0.91)	24.6	0.68 (0.96)	24.6	0.57 (0.90)

Estimates represent mean differences in HAMD scores (with corresponding posterior standard deviation) after $l = 6$ weeks. Negative values favour the first treatment. Results for Scenario 1 are based on JAGS; results for Scenario 3 are based on Stan. Note that for Trials 11, 15 and 16, no response data were available and no treatment effects could be estimated. Scenario 1 is based on participants with observed HAMD scores, whereas Scenarios 2 and 3 are based on all participants.

H_0: Trial average HAMD score at time baseline for included participants; Plac: placebo; TCA: tricyclic antidepressants; TeCA: tetracyclic antidepressants.

2.2 Scenario 2

We here consider the last observation carried forward (LOCF) approach as another common approach to account for drop-out of participants.[15,16] Advantages of the LOCF approach are that participants with missing responses are no longer excluded from the analysis (hence, preserving randomization within trials), and that its implementation is fairly straightforward.

2.3 Scenario 3

In the final scenario, we consider a more advanced approach to account for drop-out of participants and to reduce the risk of bias in estimated treatment effects. Hereto, we treat the longitudinal HAMD responses as a multivariate outcome within each trial.[17] Missing HAMD responses are then *imputed* by borrowing information from the observed HAMD responses across *all* time points within each participant. Statistical details are provided in online Appendix 2.

3 Estimation framework

In the following sections, we discuss random-effects models to combine two-arm trials for estimating the comparative efficacy of multiple treatments. For each model, we evaluate the extent of between-study heterogeneity (measured by τ^2) and the presence of network inconsistency. The presence of network inconsistency is estimated *a posteriori* by quantifying the agreement between the direct and indirect evidence for the comparison TCA: Plac (online Appendix 3). Hereto, the consistency equations of each NMA model are removed and all parameters are re-estimated. Note that the resulting (inconsistency) model can be viewed as a series of separate, independent meta-analyses for each treatment comparison, sharing a common heterogeneity variance.[1,18] This model is also known as the unrelated mean effects model. We do not present fixed effects models, as their implementation is only justifiable in the absence of heterogeneity and network inconsistency.

We here adopt a Bayesian paradigm to synthesize the treatment effects of the included trials. For all models described in this paper, we specified a uniform prior distribution U(0, 10) for the square root of the between-trial (τ^2) and within-trial (σ^2_{i6}) variance parameters.[19,20] The prior distribution for the treatment contrasts and

regression coefficients was $\mathcal{N}(0, 100^2)$. Furthermore, for the inconsistency factor we used a weakly informative prior distribution $w_{\text{TCA:Plac}} \sim U(-10, 10)$. As part of a sensitivity analysis, we informed the estimation procedure that HAMD responses are, a priori, known to lie between 0 and 54 by applying truncation: $H_{ijk6} \sim \mathcal{N}(\mu_{ijk}, \sigma^2_{i6})T(0, 54)$. All analyses were implemented in JAGS 4.0.0 (unless specified otherwise), and are based on four chains. For each chain, we allowed 50,000 adaptation samples and a burn-in period of 200,000 samples to ensure that convergence was reached (as inspected by the Gelman and Rubin diagnostic). Results are based on $4 \times 100{,}000$ samples after the burn-in period. The source code of each model is presented in online Appendix 4.

4 Meta-analysis using aggregate data

To illustrate the potential limitations of synthesizing AD, we first describe and implement meta-analysis methods that do not account for participant-level characteristics. We hereto consider the situation where only summary data (from Scenarios 1, 2 and 3) are available for the trials of our case study, and that these data have, for instance, been published in the literature.

4.1 Pairwise meta-analysis

In our case study, there are three types of trials in the simplified network of evidence: trials comparing tetracyclic antidepressants with placebo (TeCA: Plac), trials comparing tricyclic antidepressants with placebo (TCA: Plac) and trials comparing tetracyclic antidepressants with tricyclic antidepressants (TeCA:TCA). A pairwise meta-analysis (PMA) can then be achieved by implementing the usual random-effects model for each comparison:[21]

model PMA-1$^{\text{re}}$

$$\begin{aligned} \hat{y}_{i6} &\sim \mathcal{N}(\delta_i, \hat{s}^2_{i6}) \\ \delta_i &\sim \mathcal{N}\left(d_{t_i b_i}, \tau^2_{t_i b_i}\right) \end{aligned} \tag{1}$$

where b_i represents the baseline treatment and t_i represents the active treatment in trial i. The parameter $d_{t_i b_i}$ then represents the summary estimate for relative change in HAMD score between t_i and b_i. To improve the interpretation of subsequent NMA models, we also fitted a simplified PMA model with a common heterogeneity term τ^2 across all comparisons (model PMA-2$^{\text{re}}$).

In the case study, we found that treatment efficacy was largest for TCA, and lowest for placebo (Table 3). Estimated mean differences for the comparison TCA: Plac, however, varied substantially across the different scenarios. For instance, when assuming a common heterogeneity term we found that $d_{\text{TCA:Plac}}$ ranged from -2.29 (Scenario 1) to -3.91 (Scenario 2). These discrepancies are probably related to the substantial degree of drop-out, and also affected the extent of heterogeneity and inconsistency.

4.2 Network meta-analysis

Network meta-analysis is an extension of PMA where the treatment effects of different comparisons (e.g. $d_{\text{TeCA:Plac}}$ and $d_{\text{TCA:Plac}}$) are linked using a series of consistency equations. These equations state that $d_{\text{B:C}} = d_{\text{B:A}} - d_{\text{C:A}}$ for any three treatments A, B and C. Hence, when applied to our case study, we have $d_{\text{TeCA:TCA}} = d_{\text{TeCA:Plac}} - d_{\text{TCA:Plac}}$. This implies that both $\hat{d}_{\text{TeCA:TCA}}$ (direct evidence obtained from trials comparing TeCA with TCA) and $\hat{d}_{\text{TeCA:Plac}} - \hat{d}_{\text{TCA:Plac}}$ (indirect evidence obtained from trials comparing TeCA or TCA with placebo) may contribute towards inference on $d_{\text{TeCA:TCA}}$. As a result, in contrast with PMA, NMA allows to combine the evidence from *all* relevant (direct and indirect) treatment comparisons in a single statistical model:[21,22]

model NMA$^{\text{re}}$

$$\begin{aligned} \hat{y}_{i6} &\sim \mathcal{N}(\delta_i, \hat{s}^2_{i6}) \\ \delta_i &\sim \mathcal{N}(d_{t_i} - d_{b_i}, \tau^2) \quad \text{with} \quad d_1 = 0 \end{aligned} \tag{2}$$

In this model, the expression $d_{t_i} - d_{b_i}$ forms the basis of the consistency equations that combine the direct and indirect evidence for each treatment comparison. Hereby, the term d_1 is set to 0 to ensure identifiability. A common heterogeneity term τ is assumed for all treatment comparisons, for all trials.

Table 3. Results of meta-analysis models (at a follow-up time of 6 weeks) with corresponding posterior standard deviation in the simplified network of evidence.

Model	*M*	TeCA: Plac	TeCA:TCA	TCA: Plac	τ	IF	Missing data
(Network) meta-analysis using AD (Scenario 1)							
PMA-1re	15	−2.56 (1.18)	0.66 (0.76)	−2.66 (2.26)		0.57 (2.66)	M(C)AR
PMA-2re	15	−2.53 (1.17)	0.69 (0.97)	−2.29 (1.16)	1.69 (0.74)	0.93 (1.91)	M(C)AR
NMA^{re}	15	−2.18 (0.91)	0.46 (0.82)	−2.64 (0.90)	1.67 (0.66)	0.93 (1.89)	M(C)AR
NMR	15	−2.31 (0.90)	0.15 (0.74)	−2.46 (0.89)	1.26 (0.77)	1.44 (2.13)	M(C)AR
(Network) meta-analysis using AD (Scenario 2)							
PMA-1re	15	−2.03 (1.90)	0.60 (0.84)	−3.81 (1.46)		2.37 (2.54)	MAR
PMA-2re	15	−1.90 (0.92)	0.60 (0.83)	−3.91 (1.02)	1.18 (0.76)	1.41 (1.60)	MAR
NMA^{re}	15	−2.36 (0.75)	0.97 (0.71)	−3.33 (0.78)	1.21 (0.72)	1.39 (1.60)	MAR
NMR	15	−1.98 (1.01)	1.08 (0.84)	−3.06 (0.98)	1.58 (0.84)	3.07 (2.04)	MAR
(Network) meta-analysis using AD (Scenario 3)							
PMA-1^{re†}	15	−2.50 (1.64)	0.75 (0.74)	−3.57 (2.03)		0.31 (2.71)	MAR
PMA-2^{re†}	15	−2.44 (1.04)	0.75 (0.84)	−3.35 (1.09)	1.27 (0.79)	0.17 (1.72)	MAR
NMA^{re†}	15	−2.65 (0.78)	0.51 (0.71)	−3.17 (0.85)	1.17 (0.75)	1.13 (1.74)	MAR
NMR†	15	−2.83 (0.94)	0.43 (0.78)	−3.26 (0.99)	1.33 (0.87)	1.50 (2.14)	MAR
(Network) meta-analysis using IPD							
PMA-1^{ipd}	15	−2.55 (1.20)	0.67 (0.75)	−2.66 (2.25)		0.57 (2.66)	M(C)AR
PMA-2^{ipd}	15	−2.55 (1.15)	0.68 (0.96)	−2.28 (1.15)	1.65 (0.76)	0.95 (1.89)	M(C)AR
NMA^{ipd}	15	−2.17 (0.89)	0.46 (0.80)	−2.63 (0.89)	1.62 (0.68)	0.96 (1.89)	M(C)AR
NMA-PF	15	−1.99 (0.93)	0.50 (0.85)	−2.49 (0.92)	1.79 (0.72)	1.02 (1.98)	M(C)AR
NMA-TX	15	−2.81 (0.82)	0.48 (0.69)	−3.29 (0.84)	1.07 (0.72)	0.46 (1.75)	M(C)AR
MNMA†	18	−2.55 (0.52)	1.04 (0.56)	−3.59 (0.59)	0.71 (0.56)	0.73 (1.23)	MAR
(Network) meta-analysis using IPD – sensitivity analyses							
NMA^{ipd,T}	15	−3.24 (1.22)	0.98 (1.15)	−4.21 (1.28)	1.91 (1.16)	1.14 (2.73)	MAR
NMA-PF^{T}	15	−3.11 (1.17)	1.16 (1.10)	−4.27 (1.23)	1.72 (1.13)	0.81 (2.65)	MAR
NMA-TX^{T}	15	−2.76 (0.82)	0.50 (0.69)	−3.26 (0.84)	1.07 (0.72)	0.49 (1.78)	MAR
PMA^{ipd}	15	−3.16 (1.30)	0.82 (0.78)	−2.68 (2.36)		1.30 (2.80)	MNAR
NMA^{ipd}	15	−2.30 (0.89)	0.48 (0.79)	−2.77 (0.90)	1.48 (0.77)	0.73 (1.91)	MNAR
NMA-PF	15	−2.06 (0.89)	0.58 (0.78)	−2.64 (0.92)	1.44 (0.84)	0.55 (1.94)	MNAR
NMA-TX	15	−3.20 (0.79)	0.61 (0.62)	−3.20 (0.79)	0.70 (0.59)	1.14 (1.72)	MNAR

For all models, coefficients represent relative change in HAMD score after a follow-up period of 6 weeks. Negative values favour the first treatment. All estimates of mean difference and between-study heterogeneity are based on consistency models (that is, without estimation of $w_{TCA:Plac}$).
IF: Extent of network inconsistency at a follow-up time $l = 6$ weeks, given as $|w_{TCA:Plac}|$ in the inconsistency model; *M*: Number of trials used in the analyses; MNMA: multivariate NMA; NMA: network meta-analysis; NMA-PF: NMA adjusting for prognostic factors; NMA-TX: NMA allowing for treatment-covariate interaction; NMR: network meta-regression; PMA: pairwise meta-analysis.
†Corresponding models were implemented in Stan owing to the presence of multivariate missing data.
^{T}Truncation was applied to HAMD responses (range: 0 to 54).

In the case study, the NMA yielded similar results to PMA, although a substantial gain in precision was obtained for the comparative effects of TeCA and TCA versus placebo (Table 3). Although the extent of between-study heterogeneity was rather substantial for the AD generated in Scenario 1 ($\tau = 1.67$), equation (2) yielded much lower heterogeneity in Scenario 3 ($\tau = 1.17$). This confirms earlier suspicions that drop-out of participants might have affected estimates of treatment effect across trials.

Because an NMA strongly depends on the consistency of the underlying network, it is important to evaluate to what extent the direct and indirect evidence are in agreement.[18,23] We therefore specified separate models to investigate the presence of network inconsistency (online Appendix 3). In the case study, we found an inconsistency factor of $w_{TCA:Plac} = 0.93$ (Scenario 1) and $w_{TCA:Plac} = 1.13$ (Scenario 3). Although interpretation of these values is not straightforward (as the inconsistency models might yield different estimates of treatment effect and heterogeneity), they indicate substantial disagreement between the direct and indirect evidence. For instance, the inconsistency model for Scenario 3 found that the mean difference of TCA compared with placebo is less (by $w = 1.13$) when estimated indirectly via TeCA. It is possible that IPD sets were only obtained for those trials where the comparative effect of TCA versus placebo was relatively strong, for instance owing to data availability bias.[24] However, even when all relevant IPD has been obtained, it is still possible that trials

comparing TCA with placebo have been conducted in particular subgroups (where TCA is more effective). Unfortunately, there is much uncertainty around the actual extent of network inconsistency as the posterior standard deviation of $w_{\text{TCA:Plac}}$ was 1.89 (Scenario 1) and 1.74 (Scenario 3).

4.3 Network meta-regression

To improve consistency in NMA, it has been recommended that sources of variation in comparative treatment effects be explored.[25] A potential cause of heterogeneity in treatment effect is the presence of variation in effect modifiers *within* comparisons,[2,10] as this may generate different observed treatment effects in different trials. For NMA, the presence of effect modification becomes particularly problematic when there is imbalance in effect modifiers *between* comparisons, as this may lead to network inconsistency.[2]

In NMAs that are based on published AD, a meta-regression analysis can be conducted to adjust indirect comparisons for confounding bias owing to differences in the measured effect modifiers between studies. These analyses are typically based on study-level data extracted from published reports of trials. For instance, in our case study it is possible that baseline depression severity influences treatment effect. Because the trial-level averages of baseline HAMD response varied substantially across the included trials, this might have introduced heterogeneity (Table 2). We therefore consider the average baseline HAMD response in trial i, denoted $\overline{x}_i$, and centre this covariate across the included trials such that the mean of all $\overline{x}_i$ values is zero. We can then explore effect modification as follows in a random-effects network meta-regression (NMR):

model NMR

$$\begin{aligned} \hat{y}_{i6} &\sim \mathcal{N}\left(\mu_i, \hat{s}_{i6}^2\right) \\ \mu_i &= \left(\beta_{t_i} - \beta_{b_i}\right)\overline{x}_i + \delta_i \\ \delta_i &\sim \mathcal{N}\left(d_{t_i} - d_{b_i}, \tau^2\right) \quad \text{with} \quad d_1 = 0 \quad \text{and} \quad \beta_1 = 0 \end{aligned} \tag{3}$$

In Scenario 1 of the case study, we set `Plac` as the reference treatment and found $\hat{\beta}_{\text{TCA}} = -0.54$ (posterior standard deviation 0.35) and $\hat{\beta}_{\text{TeCA}} = -0.07$ (posterior standard deviation 0.46). This implies that for the comparison TeCA: Plac, trials with an average baseline HAMD response equal to the overall mean of 24.5 have a summary treatment effect of $d_{\text{TeCA:Plac}} = -2.31$. When the average baseline HAMD response in these trials decreases to 23.5, the summary treatment effect becomes

$$(-0.07 - 0) \times (23.5 - 24.5) - 2.31 = -2.24$$

Furthermore, we found that adjusting for baseline severity decreased the amount of heterogeneity from 1.67 (equation (2)) to 1.26 in Scenario 1. Conversely, when using the AD from Scenario 2 or Scenario 3, the amount of heterogeneity increased from 1.21 to 1.58, and from 1.17 to 1.33, respectively. In all scenarios, network consistency deteriorated, as $w_{\text{TCA:Plac}} = 1.44$ (Scenario 1), $w_{\text{TCA:Plac}} = 3.07$ (Scenario 2) and $w_{\text{TCA:Plac}} = 1.50$ (Scenario 3). For this reason, substantial concerns remain about the validity of estimated treatment effects.

5 Meta-analysis using individual participant data

It has previously been demonstrated that meta-regression has limited power to identify effect modification[2] and may lead to ecological bias.[13,26–28] For this reason, we here explore how the (network) meta-analysis might benefit from using IPD. We focus on one-stage meta-analysis models here, as the implementation of these models has been recommended, owing to their increased flexibility.[12,29]

5.1 Pairwise meta-analysis

The PMA model of equation (1) can be extended as follows to combine the IPD from all trials[30,31]

model PMA-1$^{\text{ipd}}$

$$\begin{aligned} H_{ijk6} &\sim \mathcal{N}\left(\mu_{ijk}, \sigma_{i6}^2\right) \\ \mu_{ijk6} &= \begin{cases} \alpha_i & : j = b \\ \alpha_i + \delta_i & : j \neq b \end{cases} \\ \delta_i &\sim \mathcal{N}\left(d_{t_i b_i}, \tau_{t_i b_i}^2\right) \end{aligned} \tag{4}$$

In this model, heterogeneity is estimated separately for each comparison and each comparison's respective treatment effect is summarized by $d_{t_i b_i}$. Although it is not required, owing to randomization, equation (4) may adjust for baseline response to increase statistical power.[12,32] Again, the term $\tau^2_{t_i b_i}$ may be replaced by a common heterogeneity term τ^2 (yielding model PMA-2$^{\text{ipd}}$) to improve the interpretation of the subsequent IPD-NMA models.

In the case study, similar estimates were obtained as for PMA based on AD (Table 3). Results did not meaningfully differ when equation (4) was adjusted for H_{ijk0} (data not shown).

5.2 Network meta-analysis

In its simplest form, a random-effects IPD-NMA can be written as

model NMA$^{\text{ipd}}$

$$\begin{aligned} H_{ijk6} &\sim \mathcal{N}\left(\mu_{ijk}, \sigma^2_{i6}\right) \\ \mu_{ijk6} &= \begin{cases} \alpha_i & : j = b \\ \alpha_i + \delta_i & : j \neq b \end{cases} \\ \delta_i &\sim \mathcal{N}\left(d_{t_i} - d_{b_i}, \tau^2\right) \quad \text{with} \quad d_1 = 0 \end{aligned} \tag{5}$$

Results in Table 3 indicate that this approach yields almost the same results as the NMA that is based on AD solely (equation (2)). Evidently, this can be expected since the AD used represent sufficient summary statistics for this particular NMA model.

5.3 Network meta-analysis adjusting for prognostic factors

As previously discussed, adjustment for baseline prognostic factors might be helpful when the meta-analysis includes trials that were poorly randomized.[33,34] It can, for instance, reduce the amount of heterogeneity in comparative treatment effects and improve overall network consistency, but also increase precision.[32,35,36] We can extend equation (5) as follows, to adjust for baseline HAMD response by treating corresponding values as a prognostic factor

model NMA-PF

$$\begin{aligned} H_{ijk6} &\sim \mathcal{N}\left(\mu_{ijk}, \sigma^2_{i6}\right) \\ \mu_{ijk6} &= \begin{cases} \alpha_i + \gamma_i x_{ijk} & : j = b \\ \alpha_i + \gamma_i x_{ijk} + \delta_i & : j \neq b \end{cases} \\ \delta_i &\sim \mathcal{N}\left(d_{t_i} - d_{b_i}, \tau^2\right) \quad \text{with} \quad d_1 = 0 \end{aligned} \tag{6}$$

In our case study, we did not find any within-trial imbalance (results not shown), and results are again very similar to equation (5) and to other meta-analysis models that are based solely on (published) AD. Furthermore, heterogeneity and network inconsistency slightly deteriorated, as we found $\tau = 1.79$ and $w_{\text{TCA:Plac}} = 1.02$, whereas these were previously (in equation (5)) estimated as $\tau = 1.62$ and $w_{\text{TCA:Plac}} = 0.96$, respectively.

5.4 Network meta-analysis adjusting for heterogeneity within and across trials

Recall that baseline HAMD response, although being balanced *within* trials, are rather imbalanced *across* trials. Furthermore, recall that we considered baseline HAMD response as a potential effect modifier of treatment effect (see section on NMR). Because meta-regression has limited power to detect effect modification and is prone to ecological bias, we here consider an IPD-NMA model that adjusts for potential effect modifiers on the participant level. Hereto, we replace the expression $\gamma_i x_{ijk}$ from equation (6) with $\gamma_{ij} x_{ijk}$ for $j = b$ and $j \neq b$. The resulting model then assumes that treatment effect is modified by x_{ijk}, and can be reformulated as follows, to separate prognostic effects from treatment–covariate interaction

model NMA-TX

$$\begin{aligned} H_{ijk6} &\sim \mathcal{N}\left(\mu_{ijk}, \sigma^2_{i6}\right) \\ \mu_{ijk6} &= \begin{cases} \alpha_i + \gamma_i x_{ijk} & : j = b \\ \alpha_i + \gamma_i x_{ijk} + \theta_i x_{ijk} + \delta_i & : j \neq b \end{cases} \\ \delta_i &\sim \mathcal{N}\left(d_{t_i} - d_{b_i}, \tau^2\right) \quad \text{with} \quad d_1 = 0 \end{aligned} \tag{7}$$

In this model, the coefficient θ_i quantifies the effect modification for treatment j due to x_{ijk} in study i. Note that equation (7) treats effect modifiers as nuisance parameters and, hence, allows the extent of effect modification to differ across studies and across comparisons. Summary estimates of effect modification can still be obtained by combining estimates of θ_i for each trial with the same comparators using fixed or random effects. Alternatively, it is possible to directly assume that θ_i is common or follows a certain distribution across studies.

In the case study, we found more favourable effects for TeCA and TCA when accounting for effect modification by baseline HAMD response. In particular, $d_{\text{TeCA:Plac}}$ changed from -2.17 (equation (5)) to -2.81 and $d_{\text{TCA:Plac}}$ changed from -2.63 (equation (5)) to -3.29. These estimates are, however, only applicable to participants with a baseline HAMD response that is equal to the average of 24. Owing to the inclusion of interaction terms, the comparative efficacy of each treatment may vary across participants. For instance, when we pooled the estimated θ_i coefficients of TeCA: Plac trials, we found an average interaction effect of $\theta = -0.29$ (standard error $= 0.20$). The proportion of total variation in the estimates of θ_i that is due to heterogeneity between studies was negligible ($I^2 = 1\%$).[37] Hence, for participants with a baseline HAMD response of 25 (instead of 24), the relative change in HAMD score between TeCA and placebo is -3.1 (instead of -2.81). Similarly, we find that the comparative effect from equation (5) ($d_{\text{TeCA:Plac}} = -2.17$) applies to those participants with a baseline HAMD response of 21.79.

5.5 Dealing with missing responses

As previously discussed, the included trials suffered from substantive drop-out, leading to a total of 31% of participants with missing responses after a follow-up period of 6 weeks (online Appendix 1). It is clear that if participants who left the trial experienced different outcomes from the remaining participants, estimated treatment effects are prone to bias.[16,38,39] This situation becomes particularly problematic in a meta-analysis, where reasons for drop-out may vary across trials. Results from our case study demonstrate that methods for dealing with missing response data (such as complete case analysis, LOCF or multivariate analyses) might yield substantially different estimates for comparative treatment effects, between-study heterogeneity and network inconsistency (Table 3). In this regard, a major advantage of IPD models is that they enable careful investigation of the influence of participants with missing responses.

In general, three missing data mechanisms can be distinguished by relating to the probability of missingness.[38,40] This probability may be independent of observed and unobserved data (missing completely at random; MCAR). Alternatively, it may be fully dependent on observed data but still independent on unobserved data (missing at random; MAR).[41] When the aforementioned IPD models are implemented without further specification, participants with missing responses are considered ignorable and do not contribute to estimation of treatment effects.[42,43] As such, analyses that exclude participants with missing responses (hence adopting MCAR) yield the same results as analyses that are based on all data (adopting MAR).[20,41] An exception to this situation occurs when HAMD responses are truncated between 0 and 54. In that case, participants with missing responses are *imputed* with informative values, and may therefore influence the posterior distribution of the meta-analysis models that are based on IPD. In the case study, we found that truncation led to stronger treatment effects but also decreased their respective precision (Table 3).

Finally, it is possible to assume that the probability of missing responses depends on both observed and unobserved data (missing not at random; MNAR).[44] The researcher must then specify an additional model that describes the pattern of missingness. For instance, in the case study it is possible that participants who are not experiencing any improvement or who are feeling worse might seek alternative treatment and drop out of the study. We can implement the following selection model to adjust for this MNAR pattern. Let m_{ijk6} represent a covariate indicating whether H_{ijk6} is observed ($m_{ijk6} = 1$) or missing ($m_{ijk6} = 0$) for participant k receiving treatment j in trial i. Then

$$m_{ijk6} \sim \text{Bernoulli}\big(p_{ijk6}\big)$$
$$\text{logit}\big(p_{ijk6}\big) = a_i + b_i\big(H_{ijk6} - H_{ijk0}\big)$$

describes that the probability of missingness depends on the participant's change in HAMD score since randomization. Note that this MNAR imputation model acknowledges the potential for heterogeneity in the drop-out mechanisms, as the regression coefficients for imputation are trial-specific.

Because effective estimation for MNAR imputation models requires mildly informative prior distributions, we assumed that participants for which HAMD scores did not improve were more likely to drop out of the study by specifying $b_i \sim \mathcal{N}(0, 10^2)T(0,)$ and $a_i \sim \mathcal{N}(0, 10^2)$. In the case study, we found that the implementation of MNAR

often decreased the extent of between-study heterogeneity and network inconsistency (Table 3). For instance, by implementing the equation from above in equation (6), the estimate for τ decreased from 1.79 to 1.44 and the extent of network inconsistency decreased from 1.02 to 0.55. However, the implementation of MNAR was not always favourable. For instance, in equation (7), we found that $\hat{w}_{\text{TCA:Plac}}$ increased from 0.46 to 1.14 despite a reduction in $\hat{\tau}$.

It is important to realize that the fit of alternative models can only be evaluated in terms of the observed data, and therefore it is impossible to verify the nature of the missingness mechanism formally.[45] For this reason, it is recommended that one conduct sensitivity analyses in which the influence of different assumptions on the achieved results is compared. It is generally preferred to improve the justification of M(C)AR analyses by specifying analysis models that adjust more extensively for the observed data. It is, for instance, possible to further extend equation (6) by incorporating HAMD responses at intermediate time points. Unfortunately, this approach becomes problematic when the intermediate responses $H_{ijk1}, \ldots, H_{ijk5}$ are not fully observed, as in the case study. We therefore consider a more elegant approach in the next section.

5.6 Modelling longitudinal responses

In the previous section, we demonstrated that drop-out is an important concern and should be carefully analysed. We therefore *imputed* missing HAMD responses by drawing from the posterior distribution of the IPD-NMA model, and evaluated several methods to specify the missing data mechanism. In this section, we go one step further and describe how the justification of MAR analyses can be improved by modelling the HAMD scores as longitudinal responses in a one-stage IPD-NMA. This approach not only enables the reduction of bias,[38,46] but may also help to increase the precision of estimated treatment effects and to include trials for which HAMD responses after 6 weeks are systematically missing (as is the case for Trials 11, 15 and 16, see Table 1).

We here propose to borrow information across time by adopting a one-stage multivariate IPD-NMA (MNMA) model.[17,47–49] In this manner, we can use information from all follow-up times across all 18 trials to estimate the treatment effects after 6 weeks. Note that in our case study, a total of 6286 HAMD scores of a possible total of 14,736 (i.e. 43%) are missing across the 18 trials. Because JAGS does not allow the specification of multivariate normal distributions with structured missing data, we use Stan 2.6.0 to estimate the following models.[50] Hereby, we used default (uniform) prior distributions for all parameters, and allowed the necessary Jacobian adjustments for variables that were declared with constraints.

We consider the following MNMA model to account for informative drop-out. This model allows for local estimates of within-study covariance[47,49] (modelled by R_i) and adopts an auto-regressive model of order one to account for potential between-study heterogeneity.

Model MNMA

$$
\begin{aligned}
&\begin{pmatrix} H_{ijk1} \\ \vdots \\ H_{ijk6} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \mu_{ijk1} \\ \vdots \\ \mu_{ijk6} \end{pmatrix}, R_i \right) \\
&\begin{pmatrix} \mu_{ijk1} \\ \vdots \\ \mu_{ijk6} \end{pmatrix} = \begin{cases} \begin{pmatrix} \alpha_{i1} \\ \vdots \\ \alpha_{i6} \end{pmatrix} & : j = b \\ \begin{pmatrix} \alpha_{i1} \\ \vdots \\ \alpha_{i6} \end{pmatrix} + \begin{pmatrix} \delta_{i1} \\ \vdots \\ \delta_{i6} \end{pmatrix} & : j \neq b \end{cases} \\
&\begin{pmatrix} \delta_{i1} \\ \vdots \\ \delta_{i6} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} d_{t_i1} - d_{b_i1} \\ \vdots \\ d_{t_i6} - d_{b_i6} \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \zeta\tau_1\tau_2 & \ldots & \zeta^5\tau_1\tau_6 \\ \zeta\tau_2\tau_1 & \tau_2^2 & \ldots & \zeta^4\tau_2\tau_6 \\ \vdots & & \ddots & \\ \zeta^5\tau_6\tau_1 & \zeta^4\tau_6\tau_2 & \ldots & \tau_6^2 \end{pmatrix} \right) \\
&R_i \sim \text{Wishart}^{-1}(\nu, \Lambda)
\end{aligned}
\tag{8}
$$

In this model, the within-study covariances, $\mathbf{R}_i$, are assumed to follow an inverse Wishart distribution with ν degrees of freedom and scale matrix $\mathbf{\Lambda}$. This parameterization for the within-study covariances enables information to be borrowed across the included trials while allowing for exchangeable correlation relationships, and therefore helps to avoid small-sample estimation fallacies. We used the uniform prior $\nu^{-1} \sim U(0,\ 0.16)$ to ensure that sampled values for the number of degrees of freedom are compatible with the dimension of the scale matrix $\mathbf{\Lambda}$.

In the case study, we found that $\zeta = 0.30$ (posterior standard deviation $= 0.74$) and $\nu = 32.50$ (posterior standard deviation $= 2.67$). Results demonstrate that equation (8) decreased the extent of between-study heterogeneity from 1.62 (equation (5)) to 0.71, and decreased the amount of network inconsistency from 0.96 to 0.73. Furthermore, when comparing the IPD analyses with the AD analyses from Scenario 3, we found that the extent of heterogeneity decreased from 1.17 (equation (2)) to 0.71 (equation (8)), and that network inconsistency decreased from 1.13 to 0.73. Finally, it is possible that further improvements can be attained by considering aforementioned treatment–covariate interactions.

6 Discussion

In this paper, we illustrated and discussed several IPD-NMA models for continuous outcome data. We explored the potential advantages of using IPD over published AD that arise when there is substantial between-study heterogeneity in treatment effect due to the presence of effect modification or missing response data. It is well known that using published AD to explore this issue might lead to ecological bias, while the use of IPD avoids this pitfall.[28] For instance, results from our case study indicate that by accounting for effect modification on the participant level (equation (7)), rather than on the trial level (equation (3)), the extent heterogeneity *and* network inconsistency considerably decreased (regardless of how the AD was generated).

Similarly, we demonstrated that access to IPD might overcome potential bias arising from informative drop-out. In particular, we found that for meta-analysis based on AD, LOCF deteriorated network consistency while complete case analysis led to excessive heterogeneity. When treating missing responses as non-ignorable by adopting a multivariate AD-NMA (equation (2) in Scenario 3) or a multivariate IPD-NMA (equation (8)), we found lower estimates of heterogeneity and network inconsistency, as compared with those models that assume non-informative drop-out (equation (2) in Scenario 1 and equation (5)). These findings confirm the need to implement appropriate methods when differential drop-out occurs,[16,38,40,46] and illustrate again the potential advantages of multivariate models.[17] Finally, we showed that even when all (published) AD is based on advanced models accounting for informative drop-out, IPD-NMA may still help to reduce the extent of between-study heterogeneity, and to improve overall precision and network consistency (as compared with NMA using AD).

Although not uniquely relevant to NMA, access to IPD also enables more rigorous sensitivity analyses. Here, we illustrated how different assumptions can be implemented when dealing with missing responses, and how statistical models can be tailored to the nature of the data. For instance, in our case study, the outcome HAMD scores represent natural numbers and are bounded between 0 and 54, and using normal approximations might not be appropriate. As we illustrated, the Bayesian Gibbs sampling framework enables the implementation of truncated response variables and the exploration of alternative missing data patterns without much additional effort. These analyses are less straightforward (and often infeasible) when operating in a frequentistic framework, and can only be implemented when IPD are at hand. The methods presented here can also be extended to analyse other outcome types, to combine published AD with IPD, to combine multi-arm studies or to include evidence from non-randomized studies.[51] Adjusting for confounders and other potential sources of bias then becomes crucial and could, for instance, be selectively applied to the relevant studies.

Some limitations need to be considered in this work. First of all, the IPD in this case study were not obtained through a comprehensive review. It is generally recommended that one combine all relevant data, and retrieve AD for trials not providing IPD. Methods and potential advantages for meta-analysing such evidence have been discussed elsewhere.[8,13] In this regard, it is important to acknowledge that the quality of the published AD might substantially affect the validity of an NMA. Results from our case study demonstrated that if trials suffer from drop-out, inclusion of AD based on complete case analysis might substantially inflate the extent of between-study heterogeneity. Unfortunately, complete case analysis is still common practice in the literature,[15] and other flawed techniques, such as LOCF,[52,53] are also likely to degrade network consistency. To explore the impact of this problem, we here considered that *all* AD is based on complete case analysis (Scenario 1), on LOCF (Scenario 2), or on multivariate analyses (Scenario 3). Evidently, published trials will often adopt a mixture of methods for dealing

with drop-out; thus, it is possible that heterogeneity and network inconsistency might further increase as a result. A second limitation is that we investigated a case study with a relatively simple network of evidence. In practice, networks tend to be more complicated and potentially include a greater diversity of treatment arms (within and across trials).[23,54] Although the presented models can be extended in a straightforward manner to combine three-armed studies, estimation of between-study covariance might become problematic in multivariate NMA models.[47,55] Finally, it is possible that the advantages found in this case study might be less (or more) pronounced in other clinical datasets. For this reason, we believe that simulation studies are needed to pin-point under what circumstances implementation of the presented models should be pursued. Furthermore, guidance is warranted to help future investigators in deciding whether the collection of IPD for all (or a certain subset of) trials would be beneficial for their particular research aim. Major known limitations of IPD include the difficulty of collecting data from all trials (thus reducing power), the possibility that trials without accessible data might differ in results from those that have accessible data (thus creating bias), and the much larger amount of time and effort to make agreements with investigators, organize data collection, clean data and perform analyses.[56–58]

In conclusion, IPD-NMA offers several potential advantages over meta-analyses that are solely based on AD. Its implementation should be considered when trials are affected by substantial drop-out, and when treatment effects are potentially influenced by participant-level covariates.

Acknowledgements

We gratefully acknowledge the constructive feedback of Jeroen Jansen during the design of this research. Finally, we are thankful for constructive feedback from the anonymous reviewers of this work.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The research leading to these results was conducted as part of the GetReal consortium. For further information please refer to www.imi-getreal.eu. This paper only reflects the personal views of the stated authors.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The work leading to these results received support from the Innovative Medicines Initiative Joint Undertaking (grant number 115546), resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/20072013) and EFPIA companies' in kind contribution. Ewoud Schuit gratefully acknowledges financial contribution for his research from the Netherlands Organisation for Scientific Research (project number 825.14.001).

Supplementary material

Supplementary material is available for this article online.

References

1. Efthimiou O, Debray TPA, van Valkenhoef G, et al. GetReal in network meta-analysis: a review of the methodology. *Res Synth Methods* 2016; **7**: 236–263.
2. Jansen JP and Naci H. Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers. *BMC Med* 2013; **11**: 159.
3. Veroniki AA, Vasiliadis HS, Higgins JP, et al. Evaluation of inconsistency in networks of interventions. *Int J Epidemiol* 2013; **42**: 332–345.
4. Song F, Xiong T, Parekh-Bhurke S, et al. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. *BMJ* 2011; **343**: d4909.
5. Song F, Altman DG, Glenny AM, et al. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *BMJ* 2003; **326**(7387): 472.
6. Dias S, Sutton AJ, Welton NJ, et al. Evidence synthesis for decision making 3: heterogeneity – subgroups, meta-regression, bias, and bias-adjustment. *Med Decis Making* 2013; **33**: 618–640.
7. White IR, Barrett JK, Jackson D, et al. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. *Res Synth Methods* 2012; **3**: 111–125.

8. Donegan S, Williamson P, D'Alessandro U, et al. 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* 2013; **32**: 914–930.
9. Salanti G, Marinho V and Higgins JPT. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. *J Clin Epidemiol* 2009; **62**: 857–864.
10. Cooper NJ, Sutton AJ, Morris D, et al. Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. *Stat Med* 2009; **28**: 1861–1881.
11. Tierney J, Vale C, Riley R, et al. Individual participant data (IPD) meta-analyses of randomised controlled trials: guidance on their use. *PLoS Med* 2015; **12**: e1001855.
12. Debray TPA, Moons KGM, van Valkenhoef G, et al. Get real in individual participant data (IPD) meta-analysis: a review of the methodology. *Res Synth Methods* 2015; **6**: 293–309.
13. Jansen JP. Network meta-analysis of individual and aggregate level data. *Res Synth Methods* 2012; **3**: 177–190.
14. Chalmers I. The Cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Ann NY Acad Sci* 1993; **703**: 156–165.
15. Bell ML, Fiero M, Horton NJ, et al. Handling missing data in RCTs; a review of the top medical journals. *BMC Med Res Methodol* 2014; **14**: 118.
16. Altman DG. Missing outcomes in randomized trials: addressing the dilemma. *Open Med* 2009; **3**: e51–e53.
17. Jackson D, Riley RD and White IR. Multivariate meta-analysis: potential and promise. *Stat Med* 2011; **30**(20): 2481–2498.
18. Dias S, Welton NJ, Sutton AJ, et al. Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomized controlled trials. *Med Decis Making* 2013; **33**: 641–56.
19. Lambert PC, Sutton AJ, Burton PR, et al. 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**: 2401–2428.
20. Gelman A, Carlin JB, Hal Stern, et al. *Bayesian data analysis*, 2nd ed. Boca Raton, FL: Chapman & Hall/CRC, 2004.
21. Dias S, Sutton AJ, Ades AE, et al. 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**: 607–617.
22. Salanti G, Higgins JPT, Ades AE, et al. Evaluation of networks of randomized trials. *Stat Methods Med Res* 2008; **17**: 279–301.
23. Higgins JPT, Jackson D, Barrett JK, et al. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Res Synth Methods* 2012; **3**: 98–110.
24. Ahmed I, Sutton AJ, Riley RD, et al. Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey. *BMJ* 2012; **344**: d7762.
25. Hutton B, Salanti G, Caldwell DM, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. *Ann Intern Med* 2015; **162**: 777–784.
26. Saramago P, Sutton AJ, Cooper NJ, et al. Mixed treatment comparisons using aggregate and individual participant level data. *Stat Med* 2012; **31**: 3516–3536.
27. Govan L, Ades AE, Weir CJ, et al. Controlling ecological bias in evidence synthesis of trials reporting on collapsed and overlapping covariate categories. *Stat Med* 2010; **29**: 1340–1356.
28. Berlin JA, Santanna J, Schmid CH, et al. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Stat Med* 2002; **21**: 371–387.
29. Debray TPA, Moons KGM, Abo-Zaid GMA, et al. Individual participant data meta-analysis for a binary outcome: one-stage or two-stage? *PLoS One* 2013; **8**: e60650.
30. Riley RD, Lambert PC, Staessen JA, et al. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Stat Med* 2008; **27**: 1870–1893.
31. Higgins JPT, Whitehead A, Turner RM, et al. Meta-analysis of continuous outcome data from individual patients. *Stat Med* 2001; **20**: 2219–2241.
32. Van Breukelen GJP. ANCOVA versus change from baseline: more power in randomized studies, more bias in nonrandomized studies [corrected]. *J Clin Epidemiol* 2006; **59**: 920–925.
33. McCarron CE, Pullenayegum EM, Thabane L, et al. Bayesian hierarchical models combining different study types and adjusting for covariate imbalances: a simulation study to assess model performance. *PLoS One* 2011; **6**: e25635.
34. Whitehead A. *Meta-analysis of controlled clinical trials*. Chichester, UK: Wiley, 2002.
35. Thompson DD, Lingsma HF, Whiteley WN, et al. Covariate adjustment had similar benefits in small and large randomized controlled trials. *J Clin Epidemiol* 2015; **68**: 1068–1075.
36. Hernández AV, Steyerberg EW and Habbema JDF. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epidemiol* 2004; **57**: 454–460.
37. Higgins JPT and Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002; **21**: 1539–1558.
38. Bell ML, Kenward MG, Fairclough DL, et al. Differential dropout and bias in randomised controlled trials: when it matters and when it may not. *BMJ* 2013; **346**: e8668.

39. Tierney JF and Stewart LA. Investigating patient exclusion bias in meta-analysis. *Int J Epidemiol* 2005; **34**: 79–87.
40. Sterne JAC, White IR, Carlin JB, et al. Quantifying heterogeneity in a meta-analysis. *BMJ* 2009; **338**: b2393.
41. Seaman S, Galati J, Jackson D, et al. What is meant by "missing at random"? *Stat Sci* 2013; **28**: 257–268.
42. Carpenter JR and Kenward MG. *Multiple imputation and its application*, 1st ed. Chichester, UK: Wiley, 2013.
43. Lesaffre E and Lawson A. *Bayesian biostatistics*. Chichester, UK: Wiley, 2012.
44. Little RJA and Rubin DB. *Statistical analysis with missing data*. Hoboken, NJ: Wiley, 2002.
45. Molenberghs G and Kenward MG. *Missing data in clinical studies*. Chichester, UK: Wiley, 2007.
46. Mallinckrodt CH, Clark WS and David SR. Accounting for dropout bias using mixed-effects models. *J Biopharm Stat* 2001; **11**: 9–21.
47. Efthimiou O, Mavridis D, Riley RD, et al. Joint synthesis of multiple correlated outcomes in networks of interventions. *Biostatistics* 2015; **16**: 84–97.
48. Trikalinos TA and Olkin I. Meta-analysis of effect sizes reported at multiple time points: a multivariate approach. *Clin Trials* 2012; **9**: 610–620.
49. Mavridis D and Salanti G. A practical introduction to multivariate meta-analysis. *Stat Methods Med Res* 2013; **22**: 133–158.
50. Stan: A C++ library for probability and sampling, Version 2.8.0, http://mc-stan.org/ (2015, accessed 3 June 2015).
51. Verde PE and Ohmann C. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. *Res Synth Methods* 2015; **6**: 45–62.
52. Lachin JM. Fallacies of last observation carried forward analyses. *Clin Trials* 2016; **13**: 161–168.
53. Dimitrakopoulou V, Efthimiou O, Leucht S, et al. Accounting for uncertainty due to 'last observation carried forward' outcome imputation in a meta-analysis model. *Stat Med* 2014; **34**: 742–752.
54. Nikolakopoulou A, Chaimani A, Veroniki AA, et al. Characteristics of networks of interventions: a description of a database of 186 published networks. *PLoS One* 2014; **9**: e86754.
55. Efthimiou O, Mavridis D, Cipriani A, et al. An approach for modelling multiple correlated outcomes in a network of interventions using odds ratios. *Stat Med* 2014; **33**: 2275–2287.
56. van Walraven C. Individual patient meta-analysis – rewards and challenges. *J Clin Epidemiol* 2010; **63**: 235–237.
57. Sud S and Douketis J. The devil is in the details… or not? A primer on individual patient data meta-analysis. *Evid Based Med* 2009; **14**: 100–101.
58. Stewart LA and Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Eval Health Prof* 2002; **25**: 76–97.