Received: 26 June 2023 | Revised: 15 December 2023 | Accepted: 18 December 2023

DOI: 10.1002/jrsm.1700

HISTORICAL REVIEW

Research Synthesis Methods WILEY

Twenty years of network meta-analysis: Continuing controversies and recent developments

A. E. Ades[1] | Nicky J. Welton[1] | Sofia Dias[2] | David M. Phillippo[1] | Deborah M. Caldwell[1]

[1]Population Health Sciences, Bristol Medical School, Bristol, UK

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

Correspondence
A. E. Ades, Population Health Sciences, Bristol Medical School, 38 Whatley Road, Bristol BS8 2PS, UK.
Email: t.ades@bristol.ac.uk

Abstract

Network meta-analysis (NMA) is an extension of pairwise meta-analysis (PMA) which combines evidence from trials on multiple treatments in connected networks. NMA delivers internally consistent estimates of relative treatment efficacy, needed for rational decision making. Over its first 20 years NMA's use has grown exponentially, with applications in both health technology assessment (HTA), primarily re-imbursement decisions and clinical guideline development, and clinical research publications. This has been a period of transition in meta-analysis, first from its roots in educational and social psychology, where large heterogeneous datasets could be explored to find effect modifiers, to smaller pairwise meta-analyses in clinical medicine on average with less than six studies. This has been followed by narrowly-focused estimation of the effects of specific treatments at specific doses in specific populations in sparse networks, where direct comparisons are unavailable or informed by only one or two studies. NMA is a powerful and well-established technique but, in spite of the exponential increase in applications, doubts about the reliability and validity of NMA persist. Here we outline the continuing controversies, and review some recent developments. We suggest that heterogeneity should be minimized, as it poses a threat to the reliability of NMA which has not been fully appreciated, perhaps because it has not been seen as a problem in PMA. More research is needed on the extent of heterogeneity and inconsistency in datasets used for decision making, on formal methods for making recommendations based on NMA, and on the further development of multi-level network meta-regression.

KEYWORDS

heterogeneity, inconsistency, medical decision making, network meta-analysis, population-adjustment, systematic review

Highlights

What is already known

- It is commonly stated that network meta-analysis relies on three assumptions (homogeneity, similarity, and consistency), that it is vulnerable to confounding like observational studies, and that direct evidence is more reliable that indirect evidence.
- A fourth assumption, transitivity, has been introduced, which requires that trials are similar in all respects apart from the treatments.
- Methodological and reporting guidelines emphasise a range of a posteriori checks on whether assumptions regarding consistency and transitivity are met.
- Heterogeneity is often tolerated or regarded as inevitable, or even desirable: to be considered in the same way as in pairwise meta-analysis.

What is new

- Exchangeability is the single assumption underlying both pairwise and network meta-analysis, but it is difficult to detect departures from it in practice, let alone verify it.
- The exchangeability assumption imposes no limit on the extent of quantitative heterogeneity.
- Heterogeneity increases the expected absolute error in pairwise comparisons, indirect comparisons, and NMA, and may introduce inconsistency between direct and indirect estimates, even when exchangeability is satisfied. This effect is due to second-order sampling variation.
- The risk of such realized error and inconsistency increases with between-studies standard deviation, and if treatment comparisons are directly informed by fewer trials.
- Most direct comparisons in treatment networks are informed by only one or two trials.

Potential impact for RSM readers

- When making recommendations based on NMA, robustness to bias, heterogeneity and inconsistency in the evidence should be checked where possible by threshold analysis. Alternatively, in clinical studies, CINeMA software can be used to reveal the impact on estimates of particular items of data.
- Every effort should be made to reduce quantitative heterogeneity, by carefully specifying the target population, by modelling and adjusting for study-related and reporting biases, by appropriately synthesizing outcomes reported in different ways and times, by avoiding treatment "lumping," and when possible by using multi-level network meta-regression to model and control for differences in patient characteristics.

1 | INTRODUCTION

Some twenty years have passed since Lumley[1] introduced the term "network meta-analysis" (NMA), referring to an extension of pairwise meta-analysis (PMA) to connected networks of randomized trial evidence, such as the network of A versus B, A versus C, A versus D, A versus E, B versus D, C versus D, illustrated in Figure 1. A 2018 bibliometric analysis[2] showed an exponential growth, recording some 2850 publications in 771 journals and in 6 languages, from over 350 institutes in 85 countries; 82% were from the USA, China, and the UK. NMA has become a core methodology in comparative effectiveness research and in health technology assessment (HTA).

Forms of network meta-analysis had appeared before Lumley's 2002 paper,[3–5] including in the 1992 Confidence Profile Method.[6] The most commonly used model is that of Higgins and Whitehead.[7] Their objective was to

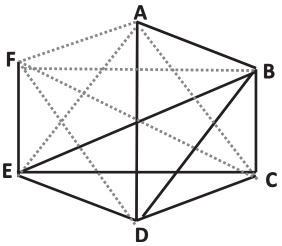

FIGURE 1 A connected network of trials on six treatments A, B, C, D, E, and F. Trial evidence (solid lines) exists on 9 of the 15 possible pairwise contrasts, AB, AD, BC, BD, BE, CD, CE, DE, EF. The remaining contrasts (dashed lines) are informed by indirect evidence. It would be possible to obtain *direct* estimates $\widehat{d}_{AB}^{Dir}, \widehat{d}_{AD}^{Dir}, \widehat{d}_{BC}^{Dir}, \widehat{d}_{BD}^{Dir} .... \widehat{d}_{EF}^{Dir}$ on each of these contrasts, but these would not be *coherent*: for example, $\widehat{d}_{BD}^{Dir} \neq \widehat{d}_{BC}^{Dir} + \widehat{d}_{CD}^{Dir}$. A *network meta-analysis* (NMA) produces coherent estimates for each of the 15 pairwise contrasts, in which $\widehat{d}_{BD}^{Coh} = \widehat{d}_{BC}^{Coh} + \widehat{d}_{CD}^{Coh}$ for every three treatments. Other terminology: *indirect* estimates can be derived from *direct* estimates as follows: $\widehat{d}_{BD}^{Ind.1} = \widehat{d}_{BC}^{Dir} + \widehat{d}_{CD}^{Dir}$ and $\widehat{d}_{BD}^{Ind.2} = \widehat{d}_{BE}^{Dir} + \widehat{d}_{ED}^{Dir}$: these two estimates involve the *evidence loops* BCD and BDE respectively.

strengthen inference on a single relative effect by introducing external indirect evidence, but their model extends readily to any number of treatments in any connected network of trials.[8–10]

The 20th anniversary of NMA is a good moment to reflect on the state of the field. In spite of its widespread and growing use, many tutorial papers, texts, and commentaries continue to express doubts about its validity and reliability.[11–15] There are also conflicting views on fundamental issues such as whether NMA evidence should be viewed as "observational,"[16] and thus whether it estimates causal effects or just associations. Further, while Lumley's original model allowed for inconsistency between direct and indirect evidence, the majority of applied work has assumed consistency, in spite of empirical evidence that inconsistency is prevalent in typical networks.[17–19] New models and parameterizations have appeared,[20,21] some in response to these findings.

The trickle of negative commentary is now impacting on patients: during stakeholder consultations on draft guidelines for treatments for depression, issued by the National Institute of Health and Care Excellence (NICE) in the UK, 23 Members of Parliament signed a motion describing NMA as "a flawed methodology,"[22] and 45 stakeholder organizations called it "an experimental technique" and pressed NICE not to base recommendations on the results from NMA.[23]

In this paper we reflect on areas of controversy that remain after the first 20 years of explosive growth in the use of NMA, suggesting possible resolutions, and reviewing new developments. We begin with a brief summary of what NMA is, and then give some context on the way NMA has been used in practice. We focus on issues of principle and interpretation, including assumptions, validity of inferences, and reliability. Some additional content appears in Supplementary Materials. We make no attempt to duplicate practical guidance already available.[16,24–31] Areas for further research are suggested throughout.

2 | WHAT IS NMA?

How should one analyse data from a connected network of trials, such as the one shown in Figure 1? The objective is to identify the best of the six treatments A, B, C, D, E, and F in a specific target population. We might consider a pairwise meta-analysis (PMA) on each of the nine sets of trials on which direct evidence is available: this would generate a series of unrelated relative treatment effect estimates $\widehat{d}_{AB}, \widehat{d}_{AC}, \widehat{d}_{AD}, \widehat{d}_{BC} .... \widehat{d}_{DE}$. However, we cannot determine the best treatment from these estimates as they lack *coherence*: in any given group of patients, and for any three treatments, the true treatment effects must obey the relationship $d_{AC} = d_{AB} + d_{BC}$, on an appropriate scale. Unrelated estimates of each pair of treatments do not have this property, which is essential for rational decision-making. NMA is a method that delivers a set of coherent estimates that have the property $\widehat{d}_{AC}^{Coh} = \widehat{d}_{AB}^{Coh} + \widehat{d}_{BC}^{Coh}$ for every set of three treatments. It also delivers coherent estimates of the additional contrasts, represented by the dashed lines in Figure 1, which have not been trialled.

A related concept which has been discussed separately is the "indirect comparison."[32] This is where inferences are made about the effect of C relative to A (which has not been compared in a trial), based on "direct" evidence from AB and BC trials, using the relationship $\widehat{d}_{AC}^{Ind} = \widehat{d}_{AB}^{Dir} + \widehat{d}_{BC}^{Dir}$. An indirect comparison is a special case of an NMA.

3 | THE FIRST TWENTY YEARS: A CHANGING CONTEXT

3.1 | Broad versus Narrow review questions and syntheses

A 2018 review in Nature noted that the same meta-analytic techniques are being used "with two different fundamental goals in mind."[33] First, the estimation of specific treatment effects in specific populations, and second the broad generalization often based on large numbers of studies, accompanied by attempts to identify the reasons for heterogeneity. In a similar vein, broad syntheses, pooling data on different products and different patient groups, have been seen as useful for testing a scientific question as to whether a class of treatment works "in principle."[34,35]

In the educational and social sciences where meta-analysis originated, the "broad" perspective has prevailed from the outset, including explicit calls to maximise between-study variation by including studies on a variety of subjects, exposed to different procedures in a range of settings.[36,37] Moderator analysis, that is sub-group analysis and meta-regression, was then recommended to identify effect modifiers.

Tutorial texts and commentaries aimed at medical audiences have also encouraged researchers to embrace a diversity of trials in their meta-analyses, declaring that heterogeneity is desirable, inevitable, or both.[24,38,39] Gøtzsche uses homoeopathy trials as an example, and clearly has the "in principle" type of analysis in mind, pointing to the loss of power if trials are split into subgroups.[40] But this argument does not work if the intention is to accurately estimate treatment effects, unless the level of heterogeneity is low. In the presence of substantial heterogeneity we can only draw conclusions about the *range* of effects.[41]

This has created a degree of ambiguity and confusion: heterogeneity is "bad" because it makes interpretation of an effect estimate difficult,[34,38] if it cannot be explained by effect modifiers, but it is also "good" because its presence allows us to explore effect modifiers.[24,37] This "apples and oranges" debate has been a mainstay of the PMA literature for many years, especially in social and psychological studies.[42]

But while social science meta-analysts could debate how to choose effect modifiers and how many could be studied,[42] their counterparts in medicine would be lucky if they could investigate even one. A study of nearly 15,000 Cochrane Reviews up to 2008 showed that 75% of meta-analyses were based on 7 or fewer studies.[43] Given that the recommended lower limit for meta-regression is 10,[44] it is surely unrealistic to believe that effect modifiers can be properly studied. Standard methods have very limited power to even detect heterogeneity,[45] let alone identify effect modifiers.[34]

3.2 | Medical decision making and systematic review

NMA is most often used in healthcare research, to determine which of a set of interventions is "best" based on all the trials that compare two or more of them. In a decision-making context, the NMA forms the evidence input for treatment recommendations. In the simplest case, the decision-maker adopts the single "best" treatment, this being the one with the highest expected value on a previously chosen evaluative criterion.[46] This could be efficacy, or cost-effectiveness based on Net Benefit,[47] which is monetized quality-adjusted life years (QALYs) minus lifetime costs, or any scheme like Multi-Criteria Decision Analysis[48] that weights different outcomes. Alternative ways of deriving treatment recommendations from an NMA are explored in Section 14.

Probabilistic[49] decision models or cost-effectiveness models, often incorporating a model of the natural history of the disease, are used by agencies that make reimbursement decisions such as the NICE,[50] and by the professional societies and colleges who issue clinical guidelines. It is essential that the joint statistical uncertainty in the NMA estimates is propagated through the decision model. This can be conveniently achieved by posterior simulation from a Bayesian NMA embedded within the decision model or, if frequentist methods are preferred, by boot-strap resampling, or by forward simulation from the maximum likelihood estimates and their variance–covariance matrix.[51] Bayesian Markov Chain Monte Carlo (MCMC) software for NMA is available from the NICE Decision Support Unit,[52] from ISPOR (International Society for Pharmaceutical Outcomes Research),[29] a 2018 textbook,[53] and several software packages.[54,55] Frequentist software is also readily available.[56,57]

Whatever the evaluative criteria, a decision context requires that interventions are narrowly defined: different doses of the same drug and different co-treatments are considered different treatments. "Lumping" over different doses or treatments is generally avoided as different doses have different costs and side-effects and are, indeed, intended to have different effects. Similarly, different treatment recommendations will be made for treatment-naïve patients and for patients who have failed on first-line therapy, so that re-imbursement decisions in particular tend to be applied to specific dose regimes in narrowly defined, clinically homogeneous populations. The decision-making context is thus inherently "narrow" in definitions of treatments and target populations.

Cochrane systematic reviews are also intended to inform clinicians' choice of treatment, but instead have emphasized completeness of inclusion of trials using the treatments under study, including grey literature. This inevitably draws together qualitatively heterogeneous sets of patients who may be at different points in the disease pathway, exemplifying the "broad" approach to synthesis. Analysts at NICE, sometimes explicitly,[58] start with a broad Cochrane review, and then select studies relevant to their narrower target population.

Interestingly, Cochrane Systematic Reviews appear to have become narrower in scope, at least in some clinical areas. The 2009 review of biologics in rheumatoid arthritis combined trials with different doses, with and without co-treatments, and in different patient groups (first line, failed on non-biologics, and failed on biologics).[59] Several years later the same authors produced separate NMAs in four different patient groups,[60–63] each review distinguishing different doses and different co-treatments.

Whether or not this represents a more general convergence in NMA practice, there remain distinct differences between NMAs undertaken for a clinical research paper and those used for decision making. This is most clearly seen in the more proactive approach to bias- and covariate-adjustment in a decision setting (Sections 12 and 13), compared to simply documenting bias[64,65] or down-grading evidence for bias, "indirectness," or other attributes.[66,67]

The first 20 years of NMA have, therefore, been a period of transition. First, a transition from the large reviews seen in the social and educational sciences, to smaller reviews in medicine, where sub-group analysis is barely feasible. Then a transition from PMA to NMA at a time when high levels of clinical and quantitative heterogeneity were either welcomed or tolerated without comment, and finally to routine use in decision making, which is generally incompatible with clinical heterogeneity, and where unexplained heterogeneity poses particular difficulties that we explore below.

4 | WHAT ARE THE ASSUMPTIONS OF NMA, AND ARE THEY DIFFERENT FROM PMA?

4.1 | Three assumptions: homogeneity, similarity, consistency

Song et al[68] described three assumptions which they stated were required by NMA. *Homogeneity*, which was also assumed in PMA, and two further assumptions, *similarity* and *consistency*, which were required for NMA. In their words:[68] homogeneity means that each trial "estimates the same single treatment effect … or different treatment effects distributed around a typical value"; similarity requires that trials in the different treatment comparison sets "are similar for moderators of relative treatment effect." Finally, consistency requires there is no conflict between the parameters estimated by "direct" and "indirect" evidence (see Figure 1). Later papers introduced another term, *transitivity*,[12,14] which requires that "indirect comparisons validly estimate the unobserved head-to-head comparison,"[12,69] equivalent to consistency. Transitivity was also seen as requiring that trials are similar in every important respect other than treatment,[70] equivalent to similarity.

Song et al's three assumptions have been repeated in methodology and tutorial papers, sometimes verbatim,[71] but also with variations such as: dropping similarity;[72] dropping homogeneity;[73] adding transitivity;[15] dropping similarity and transitivity.[74] Transitivity and similarity have been viewed as the same assumption,[12] but "similarity reduces to homogeneity" in a single head-to-head comparison.[12] Also, "the notion of transitivity is analogous to … homogeneity"; or that "a lack of transitivity causes inconsistency";[14] or that it is "incorrect" to consider transitivity and consistency as the same.[75] Evidently, there has been a lack of clarity about the precise definition of these terms, and the relationships between them. We provide a set of recommended definitions in Table 1.

4.2 | Or a single assumption: exchangeability (qualitative homogeneity)

It appears that Song et al used the term "homogeneity" as a *qualitative* construct. On this reading, the homogeneity assumption is a characterization of the standard "random effects" model, in which trial treatment effects are samples from a distribution, for example: $\delta_{i,AB} \sim Normal\left(d_{AB}, \sigma^2_{AB}\right)$. In this sense homogeneity is similar to *exchangeability* (Table 1), as previously recognized in the Bayesian literature,[76,77] although neither make specific distributional assumptions. Song et al's[68] version of the homogeneity/exchangeability assumption correctly captures the "randomness" recognized in every account of random effects meta-analysis. Consistency is then not an additional assumption required by NMA, as we ourselves once believed,[9,10] it is in fact a corollary of exchangeability.[78] The reasoning is as follows. Assuming a linear predictor scale on which treatment effects are additive, if a meta-analysis of AB trials is characterized by trial-specific relative treatment effects $\delta_{i,AB} \sim Normal\left(d_{AB}, \sigma^2_{AB}\right)$, drawn from a normal distribution with between-trial variance σ^2_{AB} and a meta-analysis of BC trials by $\delta_{i,BC} \sim Normal\left(d_{BC}, \sigma^2_{BC}\right)$, then it follows that the true AC treatment effects must conform to $\delta_{i,AC} = \delta_{i,AB} + \delta_{i,BC}$, and therefore that $\delta_{i,AC} \sim Normal\left(d_{AC}, \sigma^2_{AC}\right)$where $d_{AC} = d_{AB} + d_{BC}$ is the "consistency assumption."[10] It can also be deduced that $Minimum\left(\sigma^2_{AC}, \sigma^2_{BC}\right) < \sigma^2_{AC} \; < \sigma^2_{AB} + \sigma^2_{BC}$, which is a 2nd order consistency relating the three variances.[79] This is the triangle inequality in which the standard errors correspond to the side lengths of an acute-angled triangle (Supplementary Note 1).

Because exchangeability implies similarity with respect to effect modifiers across treatment comparisons, all three

TABLE 1 Definitions of terms.

Term	Definition
Exchangeability	Random variables are said to be exchangeable (qualitatively homogeneous), if a sequence of those variables has a joint probability distribution that is unchanged if the sequence is reordered. For example, if some studies were in treatment-naïve patients and other studies in non-naïve, then we could a priori create sequences of treatment effects that had different joint distributions.
Qualitative homogeneity/ heterogeneity	Same as Exchangeability/lack of exchangeability.
Quantitative homogeneity/ heterogeneity	The hypothesis of quantitative homogeneity is tested by, for example, Cochrane's Q-statistic,[212] or comparing the fit of fixed and random effects models. Quantitative heterogeneity is what is measured by between-trials variance. Heterogeneity refers to variation *within* treatment comparisons.
Consistency	Direct and indirect sources of evidence estimate the exact same parameters. A corollary of exchangeability.
Similarity	Distribution of effect modifiers is similar across direct and indirect sources of evidence. Implied by exchangeability.
Transitivity	(a) equivalent to consistency, or (b) equivalent to similarity
Incoherence	Same as inconsistency; lack of consistency *between* treatment comparisons.

of Song et al's requirements can be derived from exchangeability, along with transitivity in both its senses.

4.3 | Quantitative versus qualitative heterogeneity

As well as its exchangeability meaning, homogeneity has also been interpreted in a *quantitative* sense. Several authors have concluded that the homogeneity assumption requires that relative treatment effects have to be *quantitatively* similar,[71,72,80] and that it can be verified or ruled out by statistical tests of homogeneity,[27,74] or measured by I^2 statistics.[81]

However, the fact is that, whether one prefers the broad or narrow approach to synthesis, there is no technical requirement in either PMA or NMA for treatment effects to be quantitatively homogeneous; indeed, *there is no theoretical limit on how quantitatively heterogeneous treatment effects can be*. Think of a basket of green apples: however much they vary in size they are still homogeneous (exchangeable). They become non-exchangeable if, for example, oranges get in; which can happen if colour is not recognized as a potential effect modifier.

5 | HOW TO ESTABLISH EXCHANGEABILITY

If exchangeability is the only assumption, how can it be checked and verified? Establishing exchangeability a posteriori by statistical analysis requires a demonstration that subsets of the data are "similar to an adequate approximation", based on sub-group analyses, with the definitions of both "similar" and "subsets" being decided by context.[77] However, as noted above, fewer than 25% of Cochrane PMAs consist of more than seven studies,[43] and most NMAs will be under-powered to detect sub-groups. Tests for inconsistency are also investigations of whether treatment effects in pre-defined sub-sets of the data are similar. But these are also inherently weakly powered,[82] as well as being hampered by insufficient data (see Section 10).

Establishing exchangeability a posteriori will therefore seldom be feasible. Unless large numbers of trials are under study, the judgement of exchangeability of relative treatment effects in a given network can only be made a priori on the basis of topic expertise.[77] It is here that the concepts of similarity and transitivity have value, providing a rationale for a range of informal checks and investigations recommended in tutorial papers and checklists.[26,28,29,31,83] Table 2 summarises the various prior and posterior approaches to checking NMA assumptions.

6 | DOES NMA ESTIMATE CAUSAL EFFECTS OR ASSOCIATIONS?

When refereeing an applied paper recently, one of us was surprised to read the following editorial comment: "Results of this meta/network association analysis should be described in terms of association, not in terms of a causal effect."

The 2008 Cochrane Handbook[84] stated that "indirect comparisons ... are essentially observational findings across

TABLE 2 Methods for checking the exchangeability assumption, with selected references.

A posteriori checks and statistical tests	Subgroup analysis, meta-regression[213]; network meta-regression.[214]
	Single loop inconsistencies.[32]
	Multiple independent inconsistencies.[215]
	Node-Splitting.[216]
	Graphical Comparison of consistency and inconsistency models.[10]
	Design-by-treatment interaction models.[19,21,126]
	Measures of and tests for between-comparison variance.[1,10,126,217]
A priori checks for similarity/ transitivity	Examination of outcomes in control groups may provide clues about potential effect modifiers.[31,218] For example severity, previous treatment, age, calendar time.
	Examination of the distribution of potential effect modifiers across different treatment comparisons.[31,83]
	Joint randomizability:[12] it should be possible to randomize every treatment to each of the trial populations; equivalently, it should be possible to have a multi-arm trial that includes all the treatments; equivalently, in each trial treatments that are missing are missing-at-random with respect to their efficacy.

trials, and may suffer the biases of observational studies, for example due to confounding”. This statement has been repeated almost verbatim in the 2019 edition[70] and in tutorial papers on NMA.[14,27,85–87] When randomized trials are contrasted with observational studies, the term “confounding variables” usually refers to *prognostic variables*, factors that impact the absolute outcomes, but not necessarily on relative treatment effects. But indirect comparisons are *not* vulnerable to confounding in this sense, because randomization ensures that prognostic variables are balanced over arms in each trial. Thus, PMA, indirect comparisons and NMA all respect randomization and all produce weighted averages or linear combinations of relative treatment effects,[78,88] each of which are controlled for prognostic factors.

A similar claim is that NMA is a form of sub-group analysis or meta-regression, and that NMA is therefore “observational” because regression coefficients cannot be interpreted as causal effects, only as associations.[12] Both PMA and NMA, and indeed RCTs themselves, can be analysed as regressions with treatment as the covariate. But this does not prevent the coefficients being interpreted as estimates of causal effects, due to randomization.

Relative treatment effects are, however, affected by *effect modifiers*, and estimates produced by PMAs, indirect comparisons and NMAs, will be dependent on the distribution of effect modifiers in included studies.[83] In this sense PMA can be and has been regarded as observational in nature.[89,90] But, how can it be possible to draw causal inferences from a single randomized trial, but not from 2 or 3 trials?

Heterogeneity does not prevent a pooled effect from being causal, but it still has a profound impact on what can be concluded from it:

- A single trial identifies a causal effect, but in the presence of unrecognized effect modifiers, we may remain uncertain about the population or circumstances in which it applies.[38] (Possibly, this explains why trials are hard to replicate.[91])
- If, in a synthesis of several trials, the effects are quantitatively homogeneous, this increases certainty about the population in which the causal effect occurs.
- If the treatment effects are quantitatively heterogeneous, we can remain confident that there are causal effects in the populations studied, but we may now be uncertain about their size and direction in any new population.[38]
- In the extreme case of a statistically strong pooled effect, whose 95% interval does not include zero effect, but where the distribution of effects does cross the zero line, we are still confident there is a causal effect, but now we have no idea in which population(s) it occurs.

A large proportion of random effect meta-analyses appear to be in this final category (Supplementary Note 2).

7 | NETWORK GEOMETRY AND BIAS

The PRISMA-NMA check-list asks authors “to explore the geometry[92] of the treatment network … and the potential biases related to it,” and “provide an … overview of gaps in the evidence, and potential biases reflected in the network structure.” This advice requires clarification because it is unclear what kinds of bias or bias mechanisms are being associated with network structure.

Related to the idea that NMA is “observational” is that it is vulnerable to selection bias. The supposed bias occurs "when the choice of comparator in a study is dependent on the relative treatment effect."[86] A similar claim, under the name "opportunity bias," is that indirect comparisons of treatments A and C via treatment B will be biased unless allocation of patients to AB comparisons and BC comparisons is random.[93] Likewise, it has been claimed that the transitivity assumption is violated if the choice of comparators is related directly or indirectly to the relative efficacy of the interventions.[12]

Choice of comparators is far from random. New products are generally trialled against placebo to gain regulatory approval, and it is no surprise that industry trials tend—on average—to favour new drug treatments over placebo.[94] Manufacturers may also deliberately game the system by comparing their product to competitor products at a less effective dose.[95] This has earned the name "comparator preference bias".[96] But as long as the different doses are reflected as separate treatment nodes, the appropriate indirect comparisons can be made and bias caused by "lumping" different doses is avoided.

Network structure certainly needs to be monitored and understood, because it determines how potentially biased evidence in one part of the network is propagated across the network to cause bias in NMA estimates.[97,98] However, the network structure cannot lead to bias in and of itself (Supplement Note 3).

8 | IS "DIRECT" EVIDENCE BETTER THAN "INDIRECT" EVIDENCE?

A repeated claim in the literature is that "direct" evidence is better than "indirect."[28,71,99–101] Investigators have been advised to prefer direct evidence when it is available,[84] to include indirect only when direct evidence is insufficient,[17,27,87,93] to "distinguish between direct and indirect evidence", and to "justify using indirect evidence."[80] Some of these ideas may have been fuelled by the belief (Section 4) that NMA makes more assumptions, requires more checking, and is therefore inherently more dangerous and unreliable than PMA. We have not been able to find any theoretical foundation for this advice, and the standard formula[32] $\widehat{d}_{AC}^{Ind} = \widehat{d}_{AB}^{Dir} + \widehat{d}_{BC}^{Dir}$ tells us that if the direct evidence is unbiased, then the indirect estimates must be unbiased too.

The relative merits of direct and indirect evidence can be discussed if only two or three treatments are involved, but in larger networks it eventually becomes impossible to prefer one to the other or even to keep them distinct: the same evidence that is direct for one contrast is indirect for another. After all, in Figure 1, the 29 indirect estimates (19 based on 3-treatment loops and 10 on 4-treatment loops), and the set of 15 coherent NMA estimates combining both direct and indirect evidence, are all linear functions of the same 9 "direct" estimates.[78]

9 | DANGERS OF HETEROGENEITY ARE MAGNIFIED BY A SMALL NUMBER OF TRIALS

In this section we show that quantitative heterogeneity manifests itself as increased expected absolute error which affects both direct and indirect estimates and leads to inconsistency, the risk being greater as the number of trials diminishes, *even when exchangeability is satisfied*.

9.1 | Second order sampling error in PMA

Following a 1993 meta-analysis, magnesium was seen as an effective intervention following acute myocardial infarction.[102] The negative findings from the ISIS-4 super-trial therefore took the trialists by surprise,[103] while leading experts called the earlier meta-analysis results "discrepant" and "misleading."[104] However, a predictive cross-validation analysis[105,106] shows that while the treatment effect in ISIS-4 is indeed statistically significantly different from the meta-analytic pooled effect, it is entirely within the prediction interval for a new trial. In other words, given the degree of between-studies variation in previous trials, a trial with an effect as small as ISIS-4 should not have been unexpected.[53] The estimated between-study standard deviation (SD) was 0.58 on the log odds scale;[53] meaning that effects as much as a factor of 3 higher or lower than the median effect could be within the 95% envelope. This high level of heterogeneity was not commented on, or even documented, reflecting the permissive attitude to heterogeneity that prevailed at the time, and to some extent still does.

The ISIS-4 trial has been also used to illustrate publication bias,[107] and the role of sceptical Bayesian priors.[108] The confusion over whether ISIS-4 was unexpected or exactly in line with the existing evidence would be less likely to occur now, as opinion is shifting away from the pooled mean and its confidence interval as the appropriate summary, and towards the predictive effect in a new study.[109] There has also been further progress on methods for outlier detection in NMA.[110–112]

The error in the predictive effect reflects the sampling error arising from a single draw from the random effect distribution. We may extend this concept from a single study to syntheses of 2, 3 or more studies. This has been described as *second-order sampling error*.[45] Imagine a

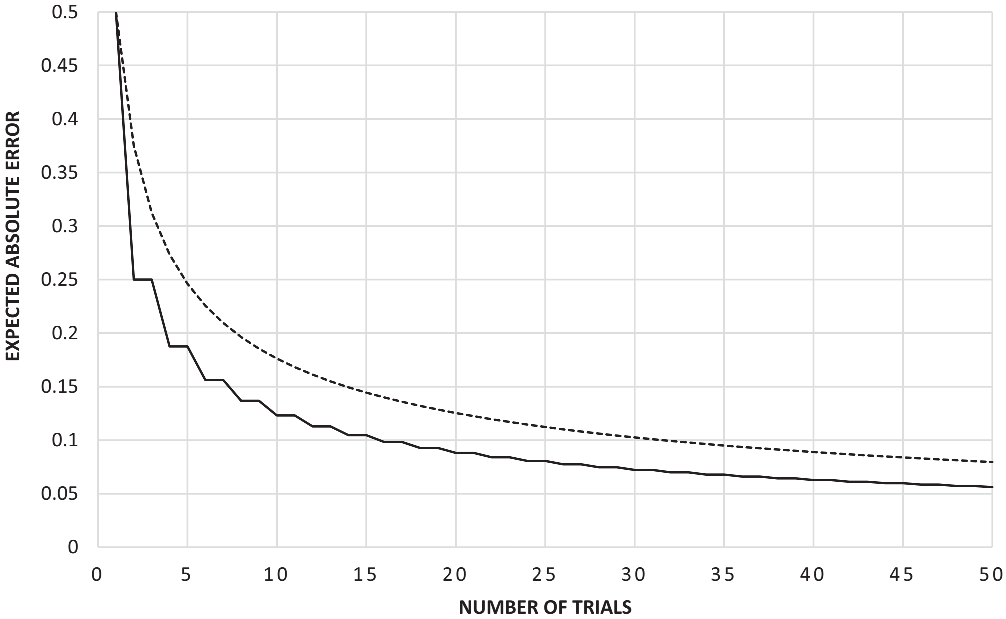

FIGURE 2 Expected absolute error in direct and indirect comparisons, in the presence of an unknown effect modifier with effect size 1 unit, present on 50% of RCTs. A is placebo, B and C are active treatments in the same class, such that the effect modifier changes their treatment effects relative to placebo to the same extent. Solid line: error in direct $\widehat{d}_{AB}^{Dir}$ estimates, and in indirect estimates $\widehat{d}_{AC}^{Ind} = \widehat{d}_{AB}^{Dir} + \widehat{d}_{BC}^{Dir}$. Dashed line: error in indirect estimates $\widehat{d}_{BC}^{Ind} = \widehat{d}_{AC}^{Dir} - \widehat{d}_{AB}^{Dir}$.

meta-analysis of infinitely large trials in which quantitative heterogeneity arises from a single unrecognized trial-level effect modifier, which is present in 50% of all trials, although this is not known to investigators. The relative treatment effect in the absence of the effect modifier is 1, and in its presence 2. The "average" treatment effect is $(0.5 \times 1.0 + 0.5 \times 2.0) = 1.5$. The expected bias of the meta-analysis relative to this target is zero, regardless of the number of trials. However, no *single* trial will ever estimate this target: it will always be 0.5 higher or 0.5 lower. If the effect modifier is known, separate analyses can be conducted on each subset. But with an unrecognized effect modifier, we do not know if a trial has delivered an overestimate or an underestimate. The expected absolute error is the appropriate statistic to reflect what we might call the *realized error* in the result of any given meta-analysis. The expected absolute error in a meta-analysis decreases as the number of trials increases,[53] from 0.5 if the meta-analysis consisted of a single trial, to 0.25 with two trials and 0.12 with 10 (see Figure 2). Recall that 75% of published Cochrane Collaboration PMAs before 2009 consisted of 7 or fewer trials.[43]

9.2 | Second order sampling error in indirect comparisons

Extending these "thought experiments" to indirect comparisons, assume that treatment A is placebo and treatments B and C are active treatments in the same class, so that the effect modifier changes the outcomes on treatments B and C in exactly the same way. The C versus B treatment effect in BC trials, $\widehat{d}_{BC}^{Dir}$, is therefore not affected by the effect modifier. In the indirect estimate: $\widehat{d}_{AC}^{Ind} = \widehat{d}_{AB}^{Dir} + \widehat{d}_{BC}^{Dir}$, the expected absolute error in $\widehat{d}_{AC}^{Ind}$ is exactly the same as the expected absolute error in the direct estimates $\widehat{d}_{AB}^{Dir}$, as shown in Figure 2. However, if we estimate the BC effect indirectly from the two active-placebo comparisons, we find that $\widehat{d}_{BC}^{Ind} = \widehat{d}_{AC}^{Dir} - \widehat{d}_{AB}^{Dir}$ now has a higher expected absolute error (Figure 2).[53] The realized error in this form of indirect comparison, which is probably the most common, is especially large.

Second order sampling error arising from heterogeneity can therefore result in realized error in direct and indirect estimates, and statistical inconsistency between them. Critically, sampling error generates inconsistency between direct and indirect estimates, *even under exchangeability*, that is even when the consistency assumption holds at the level of the parameters. The extent of the inconsistency depends only on the number of trials (Figure 2) and the degree of heterogeneity. The problem of realized error due to second order sampling only disappears if there is no heterogeneity. Analyses of many thousands of meta-analyses found median between-studies SD around 0.3 units on the standard normal scale (Supplement Note 4), which seems high when compared to the benchmark 0.2, 0.5, 0.8 for small, medium and large effect sizes.[113]

The danger of meta-analytic estimates based on small numbers of trials is especially relevant to NMAs, because most comparisons are directly informed by only one or two trials. In a report based on 201 networks published before 2019, 92% included a comparison informed by only one study; there was a median 1.3 studies per direct comparison, and the 90% percentile was less than 1.6.[82] Another collection, from 2013, of 93 Cochrane Review

NMAs had a median 1 study per comparison, and 67% had less than 2.[114] In Song et al's 2011 study of inconsistency, 50% of the 3-edged evidence loops were supported by 5 or fewer studies.[18]

It is a benefit of NMA that it allows *all* the evidence to contribute to each comparison, so that in cases where direct evidence from one or two trials delivers an extreme estimate, it is generally moderated by indirect evidence.[110] (This effect is limited by the geometry of the network; it does not hold for comparisons that involve a "dead end" in the network graph.[115]) Interestingly, in empirical studies, most of the evidence on each comparison is indirect.[116]

10 | ALTERNATIVE PARAMETERIZATIONS

10.1 | Arm-based models

The models most often used for NMA, as we noted in the introduction, are *contrast-based*,[10,117] with relative treatment effects drawn from random effect distributions. An alternative puts a multivariate-normal model on arm effects.[118–120] *Arm-based* models are equivalent to contrast-based models in which a random effect model is also put on the study-specific effects.[121] (Note that contrast-based models may have arm-based likelihoods). Putting a model on study effects has been avoided both in traditional PMA and NMA as it allows study effects to contribute information on relative treatment effects, and this risks introducing bias if the trial effect model is mis-specified.[121,122]

A second feature of arm-based models is that they oblige users to inform the absolute outcomes and the relative effects from the same data. In practice, information on absolute outcomes is better sourced from external evidence, such as register studies or single contemporary trials.[123] This is the strategy commonly adopted by decision modellers.[124]

10.2 | Inconsistency models

Two definitions of inconsistency have been proposed. *Loop inconsistency* occurs if the meta-analytic estimate of the direct AC effect, $\widehat{d}_{AC}^{Dir}$, differs from the indirect estimate $\widehat{d}_{AC}^{Ind} = \widehat{d}_{AB}^{Dir} + \widehat{d}_{BC}^{Dir}$.[10] Its detection is achieved by testing the null hypothesis that inconsistency terms for all loops $\widehat{\omega}_{ABC} = \widehat{d}_{AC}^{Dir} - \left(\widehat{d}_{AB}^{Dir} + \widehat{d}_{BC}^{Dir}\right)$ are zero.[10,125]

In networks with multi-arm trials the definition of the loop inconsistency terms depends on how the model is parameterized. *Design inconsistency*[20,126] avoids this problem by including both loop inconsistencies and differences between the AB effects estimated in AB, ABC, ABD sets of trials. It has lower power to detect inconsistency due to having more parameters than the loop-inconsistency approach.

Inconsistency models have been used primarily to detect inconsistency by goodness of fit tests and residual plots from the two models.[10,127] However, Lumley's original NMA models, which allowed for loop inconsistencies, and Jackson et al's design inconsistency models were intended for routine use in applications.[21,128] This strategy is supported by several papers reporting that inconsistency is more prevalent than would be expected by chance (Supplementary Note 5).[18,19,82]

However, the implications of these reports are unclear. First, they are based on data from Cochrane reviews and clinical literature from a period where high levels of clinical and quantitative heterogeneity were tolerated. Second, comparison of consistency and inconsistency models is limited to two methodological papers,[21,128] in which differences between estimates from consistency and inconsistency models were either barely perceptible, or differed by only a fraction of the between-studies (within-comparison) standard deviation. Model choice did not impact on the ranking of treatments by their expected effects. In practice, researchers have been hesitant to use inconsistency models in routine applications; possibly because of the dependence on parameterization of loop inconsistency models, and the difficulty understanding what design inconsistency may represent.

10.3 | Remaining research questions on inconsistency

The studies on prevalence of inconsistency should be repeated, looking separately at networks informing reimbursement decisions, clinical guideline recommendations, and other NMAs published in clinical journals. As before,[43,129] research should also distinguish between different disease areas: less between-study variation might be expected in networks of cancer treatments than in networks of depression trials or complex interventions, where trial populations are more variable. Updating these studies could provide more appropriately targeted data on suitable prior distributions for between-studies variance parameters.[43,130] Research on the prevalence of inconsistency[82] and work on inconsistency models[21] has assumed that the inconsistency terms are themselves exchangeable, drawn for example from a random effects distribution. Some of the published examples, however, suggest otherwise: plotting residuals from consistency and inconsistency models against each other can reveal that inconsistency is largely confined to just one or two comparisons.[127,131] Further research would clarify the prevalence of both exchangeable or non-exchangeable inconsistency, and the extent to which choice of model impacts on treatment recommendations.

11 | THE RELIABILITY OF NMA

A posteriori checks on exchangeability assumptions are unlikely to be conclusive (Section 5). Therefore, whatever a priori efforts are made minimize the risk of non-exchangeability, investigators need to assess the reliability of conclusions based on NMA. Distinct approaches have emerged.

11.1 | GRADE certainty of evidence ratings

GRADE is a method for ascribing a "quality" or "certainty" rating (high, moderate, low, very low) to every PMA estimate. These ratings are based on assessments in five domains: risk of bias (study limitations), imprecision, inconsistency (quantitative heterogeneity), indirectness (applicability), publication bias.[66] To extend this to NMA it was proposed that the GRADE process must be applied to every direct estimate, and then to every indirect estimate.[67] The rating attached to each indirect estimate would be the lowest of the ratings of its two direct components. Then the certainty rating attributed to each NMA estimate would be the highest of the direct and indirect ratings. In the Figure 1 network there are 21 indirect estimates based on triangular evidence loops: the same quality ratings belonging to the 9 direct estimates are therefore recycled a total of $2 \times 21 + 9 = 51$ times to produce the NMA quality ratings.

If the direct and indirect estimates are substantially different from each other, researchers are advised to choose the one with the highest certainty.[67] GRADE could then deliver an incoherent set of estimates ($A > B$, $B > C$, $C > A$). Nevertheless, the Working Group confirmed that certainty should take preference over expected efficacy[132] (though see Section 14).

The GRADE approach requires a subjective quality rating for every item of direct and indirect evidence, which is laborious to implement in networks where there may be several hundred indirect comparisons. Subsequent clarifications and elaborations[132–135] have made this time-consuming process even more complex, with unclear benefits.

11.2 | Confidence in network meta-analysis—CINeMA

The potential for incoherence and the multiple recycling of the same ratings can be avoided by exploiting the fact that each NMA estimate is a linear combination of the direct effect estimates and a set of coefficients which represent their relative contribution.[78,88] These coefficients, elements of the so-called *contributions matrix*, can be multiplied into modified GRADE ratings attaching to each direct comparison, to provide a coherent confidence rating for every NMA estimate, and a confidence rating for the whole network.[136] This approach was subsequently refined as Confidence In Network Meta-Analysis (CINeMA) and streamlined with user-friendly software[98] (Supplement Note 6).

There are drawbacks, however. First, GRADE-type assessments are required on all items of evidence at the outset, each requiring subjective judgements, including judgements of how much heterogeneity or inconsistency should be allowed before an item of evidence is downrated. As with GRADE, further complexity is introduced because of the overlap in multiple uncertainty-related concepts: imprecision, indirectness, heterogeneity, inconsistency (incoherence), and transitivity. Elaborate precautions have to be implemented to avoid double counting.[98,132,134,136]

However quality or confidence ratings for NMA estimates are generated, it is unclear how they are intended to impact on treatment decisions, because there is no relation between the confidence in evidence and the impact of that evidence on the rank ordering of treatments.[137,138]

11.3 | Threshold analysis: reliability of recommendations based on NMA

Providing an accurate analysis of the impact of every piece of direct evidence on every NMA estimate, CINeMA presents investigators with almost too much information. To reduce the dimensionality of the problem we can instead investigate the robustness of *recommendations* based on NMA evidence to potential biases and uncertainties in the data. This is done via a specific form of threshold analysis,[138,139] a standard technique in decision analysis.[46] Threshold analysis for NMA asks the question: "given the imprecision, uncertain relevance, potential biases in the trial estimates, and possible inconsistency or intransitivity, how much would the evidence have to change before this impacts the treatment recommendation?"[138,139] This question can be asked about the evidence from single trials, or the pooled evidence on each treatment contrast, or applied to treatment effects in selected subsets of trials, for example those at higher risk of bias.[140] Underlying the threshold analysis is the same algebra, based on the hat matrix, that is used in the CINeMA approach.

Threshold analysis defines the invariant intervals over which the evidence can change without impacting on recommendations. In larger NMAs it can turn out that no plausible change in the evidence, whether due to bias or sampling error, could change the recommendations.[138] (The definition of "plausible" introduces a subjective element into what is otherwise a purely mechanical process). In other cases the credible intervals extend beyond the invariant regions, indicating that the recommendation is sensitive to uncertainty.[138] Threshold analysis identifies what may be a small number of trials or treatment comparisons to which the recommendation is sensitive, so that these can be further scrutinized. By contrast, both GRADE and CINeMA require *every* item of evidence to be researched and rated at the outset, and in multiple ways. However, this disadvantage will be substantially offset by accumulating evidence in "living" reviews.[141]

Whilst GRADE and CINeMA must, in their current forms, be carried out separately for each outcome, threshold analysis can be applied to treatment recommendations based on decision models that incorporate multiple outcomes, cost-effectiveness,[139] inconsistency, and bias modelling (Section 13).[131,142] However, more work is needed to develop computational methods to apply it to the full range of non-linear economic models typically seen in HTA.

12 | POPULATION-ADJUSTED TREATMENT EFFECTS

In the presence of effect modifiers relative treatment effects estimated in trials will differ from the effects that would be observed in a target population. The purpose of population-adjustment methods is to project, or extrapolate, treatment effect estimates onto a specified target population. Solutions include: simple direct standardization;[143,144] model-based standardization,[145] a form of reweighing by propensity scores; outcome regression; and doubly robust estimation.[146] The properties of these methods, their variants, target estimands, and scope of application were reviewed previously.[147,148] Another approach reweights according to propensity calculated from a baseline risk score.[149]

Two important innovations, Matching Adjusted Indirect Comparisons (MAIC)[150] and Simulated Treatment Comparisons (STC)[151,152] extended the above methods to indirect comparisons and specifically to evidence structures including both IPD and aggregate data (AgD). MAIC and STC address a situation that is common in HTA, where, given IPD from a manufacturer's AB trial (A being placebo and B an active treatment), the aim is to generate an indirect BC comparison from AgD available from a competitor's AC trial. MAIC achieves this using a modified propensity score reweighting, and STC by covariate adjustment. Detailed analyses of both methods have been published[147,148] and their performance has been assessed in simulation studies.[153] A critical problem with both methods is that they are limited to extrapolating the AB and BC relative treatment effects into the population represented in the (competitor's) AC trial. For the same reason they can only be applied in (AB-AC) indirect comparisons, not to wider networks. We begin our description of more general methods by considering evidence networks consisting entirely of IPD.

12.1 | The relation between treatment effects at individual patient and aggregate levels

Covariate adjustment is recommended in the analysis of individual trials because it increases the precision of estimated treatment effects, controls for baseline imbalances,[154] and avoids aggregation bias when trial outcomes are non-linear.[155] For the same reasons, IPD meta-regression is recognized to be the gold standard; it also allows individual level effect modifiers to be studied and accounted for in a way that is not possible when only AgD is available.[156]

Decisions about which treatment is best are, however, made for whole populations, which requires estimation of the absolute effects of treatment at the marginal (AgD) level. Starting from an IPD meta-analysis including regression coefficients for prognostic variables, effect modifiers and treatment effects, one may obtain an average probability in a specified target population by integrating the individual outcomes over the target joint covariate distribution.[157–160] Note that for non-linear link functions, such as log, logit or probit, the relative treatment effects at the AgD and IPD levels are not the same, as has sometimes been assumed.[161–163]

12.2 | Multi-level network meta-regression (ML-NMR)

Multi-Level Network Meta-regression (ML-NMR),[153,160,164] illustrated in Figure 3, reconstructs an IPD analysis of connected networks consisting of any combination of IPD and AgD trials. It has the important property that, if all trials provide IPD, it is equivalent to the gold standard IPD meta-regression; if all provide AgD it is equivalent to a

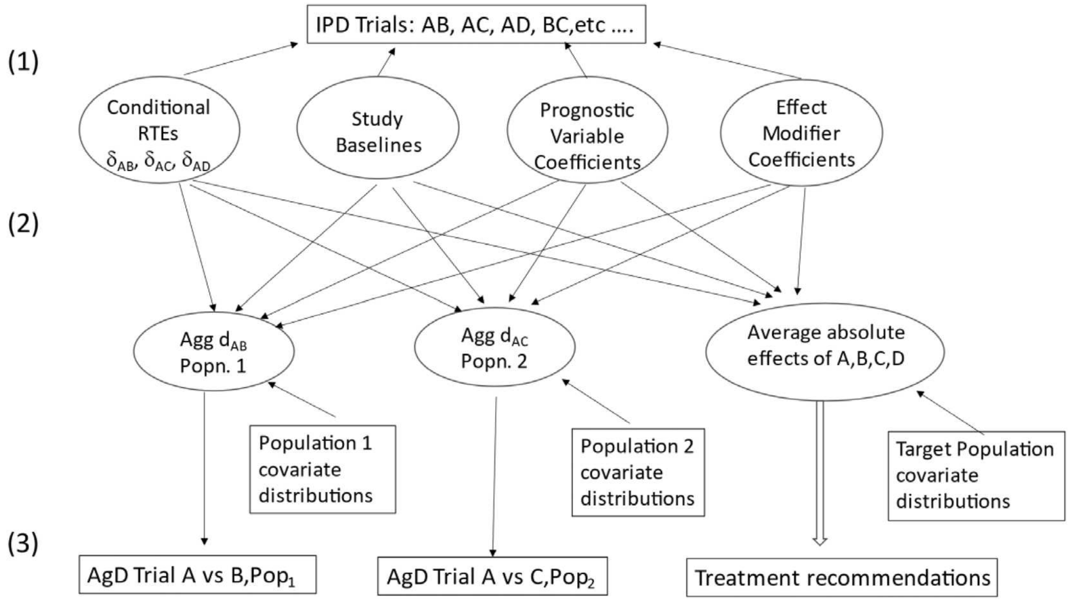
(1)
(2)
(3)
IPD Trials: AB, AC, AD, BC,etc ....
Conditional RTEs δ_AB, δ_AC, δ_AD
Study Baselines
Prognostic Variable Coefficients
Effect Modifier Coefficients
Agg d_AB Popn. 1
Agg d_AC Popn. 2
Average absolute effects of A,B,C,D
Population 1 covariate distributions
Population 2 covariate distributions
Target Population covariate distributions
AgD Trial A vs B,Pop_1
AgD Trial A vs C,Pop_2
Treatment recommendations

FIGURE 3 Multilevel network meta-regression (ML-NMR). Schematic Directed Acyclic Graph showing how (a) individual patient data, and (b) population average absolute treatment effects from aggregate data trials, jointly inform a common set of parameters, which then predict treatment effects in a target population. (1) IPD network meta-regression: relative treatment effect (RTE), study baselines, regression coefficients for prognostic variables and for effect modifiers at the IPD level are informed by a connected IPD network of trials on treatments. (2) Integration step: calculation of population average relative and absolute treatment effects in specific populations, including the target population for the decision. (3) Information from AgD trials back-propagated to contribute further parameter information.

standard NMA. ML-NMR coherently relates the individual- and aggregate-level models by integrating over the joint covariate distribution in each AgD study (Supplement Note 7), which avoids the aggregation bias caused by plugging-in mean covariate values in non-linear models, as was proposed in earlier work on combining IPD and AgD in NMA.[161–163] Unlike MAIC and STC, ML-NMR can be applied to any number of AgD and IPD trials and in networks of any complexity. Furthermore, with non-collapsible effect measures such as (log) odds ratios, ML-NMR correctly combines marginal and conditional effects and can produce both marginal and conditional estimates. Crucially, only ML-NMR can generate treatment comparisons in any specified population.

In practice, it is frequently necessary to assume that each effect modifier impacts on a set of treatments in the exact same way (the Shared Effect Modifier—SEM assumption), in order to estimate the model with the available data. In most population-adjustment analyses treatments belong to classes of drugs within which the SEM assumption is reasonable. With IPD on multiple comparisons the assumption can be tested,[164] and if necessary abandoned. Limited experience so far shows that ML-NMR estimates treatment effects more precisely than standard NMA, as it takes account of individual, within-study variation,[153,164] and reduces between-study heterogeneity and inconsistency by accounting for the different effect modifier distributions rather than averaging over them.

12.3 | Network meta-interpolation (NMI)

NMI uses information on AgD level relative treatment effects in sub-groups. If all trials report the relative treatment effect in both, say, severe and non-severe subgroups, it would be possible to estimate the relative treatment effect on the linear predictor scale in each trial that would be observed with any specified (target) proportion of severe patients. A standard NMA can then be run at the target values of that effect modifier. This can be extended to multiple sub-group dimensions using a best linear unbiased estimator approach based on marginal subgroup estimates and covariate correlations,[165] without the need for data on sub-group combinations (as with ML-NMR, the covariate correlations are borrowed from available IPD). NMI does not require the SEM assumption, as effect modifier information has to be available on all trials. However, because NMI works at the level of population average conditional effects, it cannot access the patient-level regression parameters needed to estimate absolute average treatment effects in the target population. This limits its use in HTA.

12.4 | Research priorities in population adjustment

Population adjustment methodology is an active area of research, with rapid uptake in submissions to reimbursement bodies such as NICE.[55] Several variations of MAIC and STC have been proposed,[166,167] but the fundamental inability of MAIC and STC to extend to larger networks remains, and a recent report commissioned by the NICE Decision Support Unit suggests that ML-NMR should be the preferred approach.[168] Ongoing research aims to extend ML-NMR to incorporate subgroup analyses and regression coefficients reported by AgD studies, in order to aid estimation and reduce reliance on the SEM assumption. This opens the door to performing ML-NMR analyses without any IPD at all, recreating the equivalent IPD NMA without the difficulties of obtaining IPD. Such advances have the potential to revolutionise evidence synthesis. Standardized reporting of the necessary summary statistics, including the joint covariate distributions, regression coefficients and covariance matrices, would further increase the applicability of ML-NMR. Technical methods for accessing IPD without jeopardising confidentiality or intellectual property are also being developed.[169,170]

13 | BIAS-ADJUSTMENT

13.1 | Quality-related bias in RCTs

A 1995 study by Schultz et al.[171] showed that RCTs at high risk of bias through failure of allocation concealment or lack of blinding tended to exaggerate treatment effects. Since then a series of meta-epidemiological studies have provided evidence of quality-related biases in specific disease areas and for specific kinds of outcome measure.[172–174] The bias model assumes that low risk studies estimate a relative effect δ_i and the high risk studies estimate $\delta_i + \beta_i$ with a bias distribution $\beta_i \sim N\left(b, \sigma_B^2\right)$. The meta-epidemiological studies provide empirically-based priors for bias distributions, so that studies at low and at high risk of bias can then be incorporated in a single meta-analysis.[175]

In NMA, the extra degrees of freedom can be leveraged to estimate the bias distribution associated with high-risk studies, and then generate bias-adjusted relative treatment effects, using only the data at hand. This has been applied to quality-related bias,[176] novel agent (optimism) bias,[177] small-study bias,[178] sponsor-bias,[95] and biases associated with missing data.[179,180]

13.2 | The use of non-randomized studies to estimate relative treatment effects?

The use of non-randomized studies (NRSs) to estimate relative treatment effects has been debated for over 40 years and remains controversial. Various reasons for incorporating NRS evidence, besides a lack of RCT evidence, have been advanced: that RCTs may have poor external validity,[181] that RCTs may be unable to predict treatment effects in the 'real world',[163,182] and that inclusion of NRS evidence will make the results more generalisable.[183,184] The way estimates are derived from NRS has received little attention in the synthesis literature (Supplement Note 8).

The generalisability argument takes us back to the "broad" versus "narrow" debate. If the focus is on specific treatment regimens at specific points in a disease pathway, the advantage of greater generalisability is unclear. Regarding external validity, it is true that RCT populations may have a patient mix than differs from the target distribution, but it is unclear in what way case–control and cohort studies are more "real" or externally valid than RCTs. If IPD is available, population-adjustment methods like ML-NMR (section 12) are designed to map RCT treatment effects over to any specified population, possibly based on a register, without the need to estimate relative treatment effects directly from NRS data.

A Cochrane review comparing relative treatment effects in RCTs and NRSs reported that *on average* there is little difference,[185] although "substantial heterogeneity" was reported in the odds ratio of RCT to NRS effects Thus, *in any given analysis*, a substantial absolute difference between the RCT effect and the NRS effect can be expected (Supplement Note 9). In practice, the true heterogeneity in RCT:NRS odds ratios is considerably greater, as the review examined average effects from meta-analyses consisting of between 19 and 530 studies.

13.3 | Methods for incorporating NRS evidence on relative treatment effects

Meta-analytic methods for incorporating NRS, reviewed in more detail elsewhere,[186,187] divide into three classes. In the first category NRS evidence is used to provide a prior distribution for the relative treatment effect, which can be down-weighted in various ways,[188,189] and to various extents, in recognition of the likelihood of bias.[163,181] Choosing a specific weight is problematic, which might explain why the most common approach is to accord NRSs the same weight as RCTs.[183]

A second approach introduces a third hierarchical level representing different 'types' of evidence,[163,181,182,184,190] for example RCT and NRS, perhaps also distinguishing case–control, and cohort studies. The target estimand is now the mean of the study-type distribution. But this is clearly not an estimate of the causal effect which RCTs are designed to identify. Further, its value depends on the amount of evidence on each type, and the between-type variance, which is poorly estimated with so few types, and hence influenced by priors.

Both the 3-level hierarchical model and the down-weighting methods may mitigate the bias attaching to NRS, but they inevitably incorporate it into the final estimates. The degree of influence of NRS and hence the level of bias, is subject to arbitrary choices: neither method has the transparency or consistency needed for accountable decision making.[191]

By contrast, the NMA bias-adjustment approach[176,177] (Section 13.1) can be applied to RCT and NRS evidence, to generate bias-adjusted and down-weighted estimates in a relatively objective way,[184] although modelling choices regarding direction of bias and its dependence on treatment comparison are necessary. An interesting extension[192] adds a further parameter for probability of bias. While this increases flexibility, applications so far[163,192] have used strongly informative, investigator-originated, priors for this parameter, again lacking in transparency.

14 | METHODS FOR DERIVING RECOMMENDATIONS FROM NMA

So far, we have assumed that the optimal treatment was the one with the highest expected value on the chosen evaluative criterion. This is theoretically optimal in an economic sense,[193] but the implication that uncertainty does not matter[194] jars with epidemiological thinking.

One way to liberalise the way recommendations are derived, suggested in our work on threshold analysis, would be to recommend all treatments showing an improvement of size X or more, relative to the standard reference treatment, and which are within a margin X of the most effective treatment.[138] The quantity X could, for example, be a minimal clinically important difference(MCID). Such a procedure automatically picks out the best K treatments, with K determined by the margin X. It could also be modified to take account of uncertainty by requiring a given posterior probability of superiority. A similar method was adopted in the GRADE Working Group's 2020 guidance on how to draw up recommendations.[195,196] However, the final GRADE categories are reconsidered in the light of their consistency with other pairwise comparisons, and as previously noted, a more certain treatment may in the end be preferred to a less certain one with a higher expected efficacy.[67,132]

New methods for deriving ordered treatment rankings are being actively researched.[197,198] Rankings based on the SUCRA (Surface under the cumulative ranking curve) statistic,[199] or equivalently on the P-score statistic,[200] reflect both the mean (rank) outcome value and its uncertainty. A more general formulation ranks treatments according to the probability that their effect exceeds a given threshold, and this can be extended to rankings that combine multiple outcomes, each with its own threshold.[201]

Statistical decision theory[202,203] provides another means of trading expected efficacy for certainty that is perhaps easier to justify theoretically. Put simply, the probability that a decision based on expected value is wrong gives rise to an expected loss, which can then be subtracted from the expected value of the decision, thus adjusting for uncertainty. Calculations of this sort, similar to those used to assess the value of acquiring further information,[204] have been discussed in the context of value-based pricing of pharmaceutical products.[205,206] Further research is needed to find a definition of optimality that will accommodate multiple treatment recommendations while penalising uncertainty, whether decisions are based on efficacy, cost-effectiveness, or some other evaluative function.

In the context of reimbursement decisions, methods that take account of uncertainty as well as expected value could have societal benefits, incentivizing the production of higher quality data,[207] and countering the trend to include one-arm studies and other NRS evidence in estimates of causal effects.

15 | CONCLUSIONS

15.1 | Summary of findings

The first 20 years of NMA have been a period of transition. As the PMA practices accepted in social and educational sciences were adopted in clinical studies. Investigators may have become accustomed to averaging quantitatively heterogeneous treatment effects, but with too few trials to allow causes of heterogeneity to be identified. These practices were then carried over to the newly emerging NMA. But while tutorial papers and reporting guidelines were right to cite "concerns" and "challenges,"[12,26] the emphasis was on the need to guard against inconsistency and intransitivity, while the practice of averaging over heterogeneous treatment effects remained unquestioned. In the Cochrane Handbook, the validity of NMA is said to rely on the fulfilment of the transitivity and coherence (loop consistency) assumptions, while heterogeneity is considered in the same

TABLE 3 Methods for reducing quantitative heterogeneity (between-studies variation) in evidence synthesis, with selected references.

Cause of heterogeneity	Methods to reduce heterogeneity	
Treatment variation	Variations in treatment (dose, co-treatments, and multiple components) have been averaged over in the primary analysis, creating heterogeneity, and then explored as effect modifiers rather than causal effects. Dose, treatment regime, and co-treatment should instead be considered as *treatment modifiers.* Patients cannot be randomized to different characteristics, but they can be randomized to different doses and co-treatments. NMA offers the opportunity to distinguish between different doses and co-treatments, which reduces between-study heterogeneity relative to models that "lump" treatments together. In addition, the variation in treatment can be modelled, reducing the number of parameters and increasing network connectivity.	
	Dose	For examples of dose models, see reference.[53] More complex dose models for synthesis of pharmacometric studies are also available, known as model-based network meta-analysis.[219,220]
	Treatment components	These models are used when treatments consist of multiple components, each of which has a separate effect. The effects of each component are considered to be independent and additive on the linear predictor scale, although interactions may also be modelled.[221] Examples include educational, cognitive, behavioural, and relaxation components in studies of psychological interventions;[222,223] and long-acting beta agonists and inhaled corticosteroids as components in treatments for chronic obstructive pulmonary disease.[224] Frequentist software is available.[115] These methods are likely to assume greater importance in studies of complex or behavioral interventions, where between-trials variation in treatments is particularly large.
	Class effects	Class models are used when there are several treatments that can be considered "similar,"[5,225,226] for example SSRI's for social anxiety.[227] They are a compromise between assuming that a class of different treatments all have the same effect, which may result in excessive heterogeneity, and assuming they all have independent and different effects which can lead to uncertain estimates. The class model assumes the treatment effects are drawn from a distribution.[53]
Bias Adjustment	RCT	Indicators of lower quality trials, such as lack of allocation concealment or lack of blinding, are not only associated with larger observed treatment effects (Section 13.1), but also with more between-study heterogeneity.[228] Applications of NMA bias-adjustment models[176–178] substantially reduce between-study heterogeneity by removing variation in bias. Estimation of, and adjustment for, bias, can be contrasted with approaches that simply document it or use it to downgrade evidence.[13,65,98,136]
	NRS	The same methods can be applied to Non-Randomized Studies. Much of the between-studies variance seen in NRSs is absorbed into the bias distribution, reducing the between-studies variance in relative treatment effects.[184] The other methods reviewed above (Section 13.3) result in estimates that are biased to an unknown extent.
Multiple outcomes and outcomes reported in different ways	Many trials report more than one related outcome, and different trials frequently report different outcomes. Rather than a separate synthesis for each outcome, it is preferable, especially for decision making purposes, to incorporate the different kinds of evidence into a single coherent model. The simplest approach is a within-trial synthesis combining multiple outcomes of the same type and on the same scale into a single mean effect, taking account of the correlation between them.[229] A standard univariate synthesis can then follow. Alternatively, different scores, such as verbal and maths test results, can be combined into a composite outcome.[24] For the general case of multiple outcomes, with incomplete reporting across trials, Multivariate Normal Random Effects (MVNRE) meta-analysis has been proposed and extended to network meta-analysis.[211,230–233] This is a generalization of the standard single outcome model to multivariate structures at both within- and between-trial levels, and in theory allows borrowing of strength across outcomes. Estimates from MVNRE are, however, usually very close to those from univariate analyses,[234–236] unless there is both a high proportion of missing data and a high correlation between outcomes.[237] Modelling the structural or logical relationships between outcomes directly, rather than just their correlation, allows a greater borrowing of strength across outcomes. Some examples are given below. Models of this type should be checked for clinical plausibility, and their assumptions tested statistically wherever possible.	
	Ordered categories	Outcomes in trials of treatments for Psoriatic Arthritis are often recorded in terms of the proportion showing 50%, 75%, or 90% improvement in PASI (Psoriasis Area and Severity Index). Rather than a separate synthesis at each cutoff, a more robust and inclusive analysis can be achieved by treating these as ordered categorial outcomes, with a common treatment effect at each cutoff.[53,117] This correctly captures the negative correlations between proportions of responses falling into each category, and increases network connectivity. Similar ordered benchmarks are used in Rheumatoid Arthritis trials, based on the American College of Rheumatology scale: ACR-50, ACR-75, etc.
	Scale of Measurement	Trials of treatments for depression report outcomes on a wide range of scales, such as the Beck Depression Inventory, and the Hamilton Depression Rating Scale. Patient- and Clinician-Reported Outcomes of this sort are routinely used in studies of anxiety and many other psychological and neurological disorders. The usual strategy is to convert all mean treatment effects to Standardised Mean Differences by dividing the mean effect from each study by the study standard deviation (SD). This introduces unwanted heterogeneity,[238] also known as "range variation",[45] because trial populations vary widely in their variance[239] and estimated SDs are also subject to sampling variation. One alternative is Ratio of Means,[240] which assumes that treatments act multiplicatively. If sufficient trials report on more than one scale, simultaneous synthesis and mapping of all outcomes onto a common scale has also been proposed.[239,241]

TABLE 3 (Continued)

Cause of heterogeneity		Methods to reduce heterogeneity
	Multiple time points	Binary outcomes reported at different time points can be converted to independent observations, and modelled with piece-wise constant hazards.[242,243] Multiple continuous observations in which the treatment effect is allowed to vary over time can be modelled using fractional polynomials,[244] or assuming functional forms specifically tailored to the evidence.[219,220]
	Structural relationships	Cancer studies may report either time to progression-free survival (PFS) or overall survival (OS), or both. A common approach has been to model the OS and PFS log hazard ratios using MVNRE,[245] and similar models are used routinely in the surrogate endpoint literature.[246–248] An alternative approach is to estimate NMA models of PFS and Post-Progression Survival (PPS), subject to the structural constraint PSF + PPS=OS, using area-under- the-curve (AUC) as an outcome, up to a restricted follow-up time.[249] This method requires that Kaplan Meier curves are digitized so that AUC can be measured.[250] The limitation to restricted mean survival can be avoided by using external register data to extrapolate survival curves.[251,252] Other examples where structurally related outcomes are reported in a variety of ways are: (1) combining data on PFS, OS and probability of response in a single model.[253] (2) combining data on median survival time, mean survival time, and proportion surviving.[254] (3) 'Chain of evidence' structures in which trials report one or more of: the proportion of patients reaching endpoint A, the proportion reaching later endpoint B, and the proportion reaching B conditional on having reached A.[255]
Survival models		Synthesis of time-to-event (survival) trials is usually based on the reported hazard ratios from Cox survival models, assuming proportional hazards (PH). Most trial reports include Kaplan Meier curves and algorithms have been published which reconstruct the original curve from the digitised image.[250] Although abandonment of the PH assumption requires investigators to adopt new measures of treatment effect,[256] it allows a wide range of more flexible models,[257,258] including 2- and 3-parameter survival curves,[257,259] fractional polynomials,[260] restricted mean survival time,[261] splines,[262] and piece-wise exponential models.[242]

way as in PMA.[70] Similarly, AHRQ guidance calls for (quantitative) homogeneity of direct evidence in NMA, but notes that this applies to all forms of synthesis, and the emphasis is again on checking inconsistency.[27] The fact that heterogeneity can itself introduce inconsistency, especially when there are few trials, and the dire impact, of having most NMA comparisons informed directly by only one or two trials (Figure 2), has not been widely appreciated.

NMA finds its natural place in medical decision making, where it is needed to compare specific treatments in specific patient populations. In this context, where quantitative heterogeneity and inconsistency are likely to be far less, NMA may be less problematic.

This paper has tried to clarify, if not resolve, the continuing controversies surrounding NMA. Our findings are:

- There is only one assumption, exchangeability, from which the consistency assumption can be derived, and which implies similarity and transitivity. However, this assumption is weak and hard to test, let alone verify, especially in small datasets; it is only in larger NMAs that it has any practical significance.
- Exchangeability places no limit on quantitative heterogeneity. But heterogeneity results in lack of interpretability, and when contrasts are informed by few trials, can increase realized error and lead to inconsistency between direct and indirect estimates.
- Estimates of relative treatment effects from PMA, indirect comparisons and NMA are estimates of causal effects: any interactions observed in investigations of subgroups are associations. For this reason, it is important that differences in treatments, such as dose differences, are analysed as different treatments, not as sub-groups.
- Network geometry is important to understanding how biased evidence in one part of the network could influence estimates elsewhere. It may also reflect biases in what evidence is collected or reported, but it cannot in itself cause NMA estimates to be biased.

15.2 | Recommendations for practice

Besides general recommendations on the need for threshold analyses, and other forms of sensitivity analysis such as quantitative bias analysis,[208] our main recommendation, in line with our main findings, is that quantitative heterogeneity should be minimized. Heterogeneity undermines the relevance of estimates to the target population (Section 6), and increases realized error and inconsistency, even under exchangeability, especially when contrasts are informed by small numbers of studies (Section 9). Reducing heterogeneity will also reduce the impact of any lack of exchangeability.

Clinical heterogeneity is an important source of between-trials variation. However, the decision to adopt either a "narrow" or "broad" approach to patient inclusion is best taken by re-imbursement agencies or guideline developers, not methodologists, while bearing in mind that broad inclusion may result in quantitative heterogeneity and put recommendations at increased risk of error, to an extent determined by the between-studies SD. It is concerning that it may be difficult to reduce the heterogeneity in studies of psychological or complex interventions, where there is so much variation between interventions and their implementation as well as variation in study populations (Supplement Note 10).

There is a growing body of literature on statistical modelling methods for reducing quantitative heterogeneity in evidence synthesis (Table 3). This includes variation in treatments (dose, treatment components), bias-adjustment, methodological variation, and time-to-event (survival) outcomes. Many of these strategies are under-utilized.[209] Also included in Table 3 are a range of multivariate methods that allow disparate forms of data, that would otherwise be analysed separately, to be aggregated into a single coherent analysis. This increases network connectivity and robustness of conclusions. Separate analysis of multiple outcomes ignores within-study correlations, increases heterogeneity, lowers the precision of estimates,[210] and creates a multiplicity problem.[211] ML-NMR and NMI should be adopted whenever possible: these are the key methodologies capable of reducing the impact of between-trials variation in patient characteristics.

AUTHOR CONTRIBUTIONS

A. E. Ades: Conceptualization; methodology; writing – original draft; writing – review and editing. **Nicky J. Welton:** Methodology; writing – review and editing. **Sofia Dias:** Methodology; writing – review and editing. **David M. Phillippo:** Methodology; writing – review and editing. **Deborah M. Caldwell:** Conceptualization; methodology; writing – review and editing.

ACKNOWLEDGMENTS

The authors would like to thank JRSM's editors and reviewers for their contributions to the paper.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analysed in this study.

ORCID

A. E. Ades https://orcid.org/0000-0001-7822-3552
Nicky J. Welton https://orcid.org/0000-0003-2198-3205
Sofia Dias https://orcid.org/0000-0002-2172-0221
David M. Phillippo https://orcid.org/0000-0003-2672-7841

REFERENCES

1. Lumley T. Network meta-analysis for indirect treatment comparisons. *Stat Med*. 2002;21(16):2313-2324.
2. Shi J, Gao Y, Ming L, et al. A bibliometric analysis of global research output on network meta-analysis. *BMC Med Inform Decising Making*. 2021;21(144).
3. Gleser LJ, Olkin I. Stochastically dependent effect sizes. In: Cooper H, Hedges LV, eds. *The Handbook of Research Synthesis*. Russell Sage Foundation; 1994:339-355.
4. Hasselblad V. Meta-analysis of multi-treatment studies. *Med Decis Making*. 1998;18:37-43.
5. Dominici F, Parmigiani G, Wolpert RL, Hasselblad V. Meta-analysis of migraine headache treatments: combining information from heterogenous designs. *J Am Stat Assoc*. 1999;94:16-28.
6. Eddy DM, Hasselblad V, Shachter R. *Meta-Analysis by the Confidence Profile Method*. Academic Press; 1992.
7. Higgins JPT, Whitehead A. Borrowing strength from external trials in a meta-analysis. *Stat Med*. 1996;15:2733-2749.
8. Lu G, Ades A. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med*. 2004;23:3105-3124.
9. Caldwell DM, Ades AE, Higgins JPT. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ*. 2005;331:897-900.
10. Lu G, Ades AE. Assessing evidence consistency in mixed treatment comparisons. *J Am Stat Assoc*. 2006;101:447-459.
11. Gartlehner G, Moore C. Direct versus indirect comparisons: a summary of the evidence. *Int J Technol Assess Health Care*. 2008;24(2):170-177.
12. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Res Synth Methods*. 2012;3:80-97.
13. Li T, Puhan MA, Vedula SS, Singh S, Dickersin K. Network meta-analysis: highly attractive but more methodological research is needed. *BMC Med*. 2011;9:79.
14. Cipriani A, Higgins JPT, Geddes JR, Salanti G. Conceptual and technical challenges in network meta-analysis. *Ann Intern Med*. 2013;159(2):130-137.
15. Faltinsen EG, Storebø OJ, Jakobsen JC, Boesen K, Lange T. Network meta-analysis: the highest level of medical evidence? *EBM Anal*. 2018;23:56-59.
16. Higgins J, Thomas J, Chandler J, et al., eds. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd ed. Wiley; 2019.
17. Song F, Altman D, Glenny A-M, Deeks J. Validity of indirect comparison for estimating efficacy of competing interventions: evidence from published meta-analyses. *Br Med J*. 2003;326:472-476.
18. 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.
19. Veroniki AA, Vasiliadis HS, Higgins JPT, Salanti G. Evaluation of inconsistency in networks of interventions. *Int J Epidemiol*. 2013;42:332-345.

20. Higgins JP, Jackson D, Barrett JK, Lu G, Ades AE, White IR. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Res Synth Methods*. 2012;3(2):98-110.
21. Jackson D, Barrett JK, Rice S, White IR, Higgins JPT. A design-by-treatment interaction model for network meta-analysis with random inconsistency effects. *Stat Med*. 2014;33:3639-3654.
22. UK Parliament. Early Day Motion 980: NICE Guidelines on Depression. 2018 https://edm.parliament.uk/early-day-motion/51449/nice-guidelines-on-depression-in-adults
23. National Institute for Health and Care Excellence. Stakeholder position statement on the NICE guideline for depression in adults, July. 2021 https://www.bacp.co.uk/media/13407/stakeholder-position-statement-on-the-nice-guideline-for-depression-in-adults.pdf, https://cdn.ymaws.com/www.psychotherapyresearch.org/resource/resmgr/uk-spr/spruk_nice_01_2022.pdf
24. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. *Introduction to Meta-Analysis*. Wiley; 2009.
25. Cooper H, Hedges LV, Valentine JC. *The Handbook of Research Synthesis and Meta-Analysis*. 2nd ed. Russel Sage Foundation; 2009.
26. 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(11):777-784.
27. Morton SC, Murad MH, O'Connor E, et al. *Quantitative Synthesis—An Update. Methods Guide for Comparative Effectiveness Reviews*. Agency for Healthcare Quality and Research; 2018.
28. Jansen JP, Fleurence R, Devine B, et al. Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ISPOR task force on indirect treatment comparisons good research practices: part 1. *Value Health*. 2011;14:417-428.
29. Hoaglin DC, Hawkins N, Jansen JP, et al. Conducting indirect-treatment-comparison and network-meta-analysis studies: report of the ISPOR task force on indirect treatment comparisons good research practices: part 2. *Value Health*. 2011;14(4):429-437.
30. Higgins JPT, Lasserson T, Thomas J, Flemyng E, Churchill R. Methodological expectations of cochrane intervention reviews (MECIR). London. 2023. https://community.cochrane.org/mecir-manual
31. Cope S, Zhang J, Saletan S, Smiechowski B, Jansen JP, Schmid P. A process for assessing the feasibility of a network meta-analysis: a case study of everolimus in combination with hormonal therapy versus chemotherapy for advanced breast cancer. *BMC Med*. 2014;12(93).
32. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol*. 1997;50(6):683-691.
33. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. *Nature*. 2018;555:175-182.
34. Thomas J, Kneale D, McKenzie JE, Brennan SE, Bhaumik S. Determining the scope of the review question and the question it will address. In: Higgins JPT, Thomas J, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd ed. Wiley; 2019.
35. Caldwell DM, Welton NJ. Approaches for synthesising complex mental health interventions in meta-analysis. *Evid Based Mental Health*. 2015;19(1):16-21.
36. Cooper H. Hypotheses and problems in research synthesis. In: Cooper H, Hedges LV, Valentine JC, eds. *The Handbook of Research Synthesis and Meta-Analysis*. 2nd ed. Russell Sage Foundation; 2009.
37. Cooper H, Hedges LV. Potentials and limitations. In: Cooper H, Hedges LV, Valentine JC, eds. *The Handbook of Research Synthesis and Meta-Analysis*. 2nd ed. Russell Sage Foundation; 2009.
38. Thompson SG. Why sources of heterogeneity in meta-analyses should be investigated. *Br Med J*. 1994;309:1351-1355.
39. Higgins J, Thompson S, Deeks J, Altman DG. Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. *J Health Serv Res Policy*. 2002;7(1):51-61.
40. Gøtzsche PC. Why we need a broad perspective on meta-analysis. *Br Med J*. 2000;321:585-586.
41. Borenstein M. *Common Mistakes in meta-Analysis and how to Avoid them*. Biostat Inc; 2019.
42. Sharpe D. Of apples and oranges, file drawers and garbage: why validity issues in meta-analysis will not go away. *Clin Psychol Rev*. 1997;17(8):881-901.
43. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JPT. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane database of systematic reviews. *Int J Epidemiol*. 2012;41:818-827.
44. Deeks J, Higgins JPT, Altman DG. Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd ed. Wiley; 2019.
45. Hunter JE, Schmidt FL. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. 2nd ed. Sage Publications; 2004.
46. Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. *Methods for the Economic Evaluation of Health Care Programmes*. 4th ed. Oxford University Press; 2015.
47. Stinnett A, Mullahy J. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analyses. *Med Decis Making*. 1998;18:S68-S80.
48. Thokala P, Duenas A. Multiple criteria decision analysis for health technology assessment. *Value Health*. 2012;15:1172-1181.
49. Doubilet P, Begg CB, Weinstein MC, Braun P, McNeill BJ. Probabilistic sensitivity analysis using Monte Carlo simulation: a practical approach. *Med Decis Making*. 1985;5:157-177.
50. National Institute for Health and Care Excellence. Guide to the methods of technology appraisal. London. 2013. https://www.nice.org.uk/process/pmg9
51. Dias S, Sutton AJ, Welton NJ, Ades AE. Evidence synthesis for decision making 6: embedding evidence synthesis in probabilistic cost-effectiveness analysis. *Med Decis Making*. 2013;33:671-678.
52. Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical Support Document 2: A generalised linear modelling framework for pair-wise and network meta-analysis of randomised controlled trials. 2011.
53. Dias S, Ades AE, Welton NJ, Jansen JP, Sutton AJ. *Network Meta-Analysis for Decision Making*. Wiley; 2018.

54. van Valkenhoef G, Lu G, de Brock B, Hillege H, Ades AE, Welton NJ. Automating network meta-analysis. *Res Synth Methods*. 2012;3(4):285-299.

55. Phillippo DM. multinma: Bayesian network meta-analysis of individual and aggregate data. R-package version 0.5.1, *Zenodo*, 2023. https://CRAN.R-project.org/package=multinma or 10.5281/zenodo.3904454

56. White IR. Multivariate random-effects meta-regression: updates to mvmeta. *Stata J*. 2011;11:255-270.

57. Rücker G, Krahn U, König J, et al. netmeta: Network meta-analysis using frequentist methods. R package Version 2.8–2, R Development Core Team. 2023. https://cran.r-project.org/web/packages/netmeta/index.html

58. National Institute for Health and Care Excellence. Vaccine Uptake in the General Population (NG213). London. 2022.

59. Singh JA, Christensen R, Wells GA, et al. Biologics for rheumatoid arthritis: an overview of Cochrane reviews. *Cochrane Database Syst Rev*. 2009;2009(4).

60. Singh JA, Hossain A, Mudano AS, et al. Biologics or tofacitinib for people with rheumatoid arthritis naive to methotrexate: a systematic review and network metaanalysis. *Cochrane Database Syst Rev*. 2017;5.

61. Singh JA, Hossain A, Tanjong Ghogomu E, Mudano AS, Tugwell P, Wells GA. Biologic or tofacitinib monotherapy for rheumatoid arthritis in people with traditional disease-modifying anti-rheumatic drug (DMARD) failure: a Cochrane systematic review and network meta-analysis (NMA). *Cochrane Database Syst Rev*. 2016;11.

62. Singh JA, Hossain A, Tanjong Ghogomu E, et al. Biologics or tofacitinib for rheumatoid arthritis in incomplete responders to methotrexate or other traditional disease-modifying anti-rheumatic drugs: a systematic review and network meta-analysis. *Cochrane Database Syst Rev*. 2016;5.

63. Singh JA, Hossain A, Tanjong Ghogomu E, et al. Biologics or tofacitinib for people with rheumatoid arthritis unsuccessfully treated with biologics: a systematic review and network meta-analysis. *Cochrane Database Syst Rev*. 2017;2017:3.

64. Sterne JAC, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016;355:1-7.

65. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *Br Med J*. 2019;366:14898.

66. Guyatt G, Oxman A, Vist G, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336:924-926.

67. Puhan MA, Schünemann HJ, Murad MH, et al. A GRADE working group approach for rating the quality of treatment effect estimates from network meta-analysis. *BMJ*. 2014;349:g5630.

68. Song F, Loke Y-K, Walsh T, Glenny A-M, Eastwood AJ, Altman DG. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. *Br Med J*. 2009;338(31):b1147.

69. Barbui C, Cipriani A, Furukawa TA, et al. Making the best use of available evidence: the case for new generation antidepressants. *Evid Based Mental Health*. 2009;12:101-104.

70. Chaimani A, Caldwell DM, Li T, Higgins JPT, Salanti G. Undertaking network meta-analyses. In: Higgins JPT, Thomas J, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd ed. Wiley; 2019.

71. Mills EJ, Ioannidis JPA, Thorlund K, Schunemann HJ, Puhan MA, Guyatt GH. How to use an article reporting a multiple treatment meta-anlysis. *JAMA*. 2012;308:1246-1253.

72. Watt J, Tricco A, Straus S, Veroniki A, Naglie G, Drucker A. Research techniques made simple: network meta-analysis. *J Invest Dermatol*. 2019;139:4-12.

73. Kanters S, Ford N, Druyts E, Thorlund K, Mills E, Bansback N. Use of network meta-analysis in clinical guidelines. *Bull World Health Organ*. 2016;94:782-784.

74. Donegan S, Williamson P, D'Alessandro U, Tudur SC. Assessing key assumptions of network meta-analysis: a review of methods. *Res Synth Methods*. 2013;4(4):291-323.

75. Fernandez-Castilla B, Van den Noortgate W. Network meta-analysis in psychology and educational sciences: a systematic review of their characteristics. *Behav Res Methods*. 2023;55:2093-2108.

76. Barnardo JM. Modern Bayesian inference: foundations and objective methods. In: Bandyupadhyay PS, Forster MR, eds. *Philosophy of Statistics*. North Holland; 2011.

77. Draper D, Hodges JS, Mallows CL, Pregibon D. Exchangeability and data analysis. *J Royal Stat Soc Series A, (Stat Soc)*. 1993;156(1):9-28.

78. Lu G, Welton NJ, Higgins JPT, White IR, Ades AE. Linear inference for mixed treatment comparison meta-analysis: a two-stage approach. *Res Synth Methods*. 2011;2(1):43-60.

79. Lu G, Ades AE. Modelling between-trial variance structure in mixed treatment comparisons. *Biostatistics*. 2009;10:792-805.

80. Donegan S, Williamson P, Gamble C, Tudor-Smith C. Indirect comparisons: a review of reporting and methodological quality. *PloS One*. 2011;5:e11054.

81. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21:1539-1558.

82. Veroniki AA, Tsokani S, White IR, et al. Prevalence of evidence of inconsistency and its association with network structural characteristics in 201 published networks of interventions. *BMC Med Res Methodol*. 2021;21(224):224.

83. Jansen J, 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.

84. Higgins JPT, Green S. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.0 [Updated February 2008]*. The Cochrane Collaboration, Wiley; 2008.

85. Becker LA, Oxman AD. Chapter 22: overviews of reviews. In: Higgins J, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions Version 500 (Updated February 2008)*. The Cochrane Collaboration; 2008.

86. Efthimiou O, Debray TPA, van Valkenhoef G, et al. GetReal in network meta-analysis: a review of the methodology. *Res Synth Methods*. 2016;7(3):236-263.

87. ter Veer E, van Oijen MGH, van Laarhoven HWM. The use of (network) meta-analysis in clinical oncology. *Front Oncol*. 2019;9:822.

88. Krahn U, Binder H, Konig J. A graphical tool for locating inconsistency in network meta-analyses. *BMC Med Res Methodol*. 2013;13:35.

89. Victor N. Indications and contra-indications for meta-analysis. *J Clin Epidemiol*. 1995;48:5-8.

90. Egger M, Smith GD. Meta-analysis. Potentials and promise. *Br Med J*. 1997;315:1371-1374.
91. Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.
92. Salanti G, Kavvoura FK, Ioannidis JPA. Exploring the geometry of treatment networks. *Ann Intern Med*. 2008;148:544-553.
93. Edwards SJ, Clarke MJ, Wordsworth S, Borrill J. Indirect comparisons of treatments based on systematic reviews of randomised controlled trials. *Int J Clin Pract*. 2009;63(6):841-854.
94. Dias S, Welton NJ, Ades AE. Study designs to detect sponsorship and other biases in systematic reviews. *J Clin Epidemiol*. 2010;63:587-588.
95. Naci H, Dias S, Ades AE. Industry sponsorship bias in research findings: a network meta-analytic exploration of LDL cholesterol reduction in the randomised trials of statins. *BMJ*. 2014;349:g5741.
96. Ioannidis JPA. Integration of evidence from multiple meta-analyses: a primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. *Can Med Assoc J*. 2009;181:488-493.
97. Li H, Shih M-C, Song C-J, Tu Y-K. Bias propagation in network meta-analysis models. *Res Synth Methods*. 2023;14:247-265.
98. Nikolakopoulou A, Higgins JPT, Papakonstantinou T, et al. CINeMA: an approach for assessing confidence in the results of a network meta-analysis. *PLoS Med*. 2020;17(4):e1003082.
99. Ioannidis JPA. Indirect comparisons: the mesh and mess of clinical trials. *Lancet*. 2006;368:1470-1472.
100. *World Health Organisation*. WHO Handbook for Guideline Development; 2012.
101. Deeks JJ, Altman DG, Bradburn MJ. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In: Egger M, Davey Smith G, Altman DG, eds. *Systematic Reviews in Health Care: Meta-Analysis in Context*. BMJ Publishing Group; 2004:285-312.
102. Yusuf S, Koon T, Woods K. An effective, safe, simple, and inexpensive intervention. *Circulation*. 1993;87:2043-2046.
103. Yusuf S, Flather M. Magnesium in acute myocardial infarction. *Br Med J*. 1995;301(751):751-752.
104. Egger M, Davey-Smith G. Misleading meta-analysis. *Br Med J*. 1995;310:752-754.
105. DuMouchel W. Predictive cross-validation of Bayesian meta-analyses. In: Bernardo JM, Berger JO, Dawid AP, Smith AFM, eds. *Bayesian Statistics 5*. Oxford University Press; 1996:107-127.
106. Marshall EC, Spiegelhalter DJ. Approximate cross-validatory predictive checks in disease mapping models. *StatMed*. 2003;22(10):1649-1660.
107. Egger M, Davey SG, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *Br Med J*. 1997;315:629-634.
108. Higgins JPT, Spiegelhalter DJ. Being sceptical about meta-analyses: a Bayesian perspective on magnesium trials in myocardial infarction. *Int J Epidemiol*. 2002;31(1):96-104.
109. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Res Synth Methods*. 2017;8(1):5-18.
110. Madan J, Stevenson MD, Ades AE, Cooper KL, Whyte S, Akehurst R. Consistency between direct and indirect trial evidence: is direct evidence always more reliable? *Value Health*. 2011;14:953-960.
111. Petropoulou M, Salanti G, Rucker G, Schwarzer G, Moustaki I, Mavridis D. A forward search algorithm for detecting extreme study effects in network meta-analysis. *Stat Med*. 2021;40:5642-5656.
112. Metelli S, Mavridis D, Créquit P, Chaimani A. Bayesian model-based outlier detection in network meta-analysis. *J Roy Stat Soc A (Stat Soc)*. 2023;184:754-771.
113. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press; 1969.
114. Xiong T, Parekh-Bhurke S, Loke YK, et al. Overall similarity and consistency assessment scores are not sufficiently accurate for predicting discrepancy between direct and indirect comparison estimates. *J Clin Epidemiol*. 2013;66:184-191.
115. Rücker G, Petrolpoulou M, Schwarzer G. Network meta-analysis of multicomponent interventions. *Biom J*. 2020;62(3):808-821.
116. Papakonstantinou T, Niokolakopoulou A, Egger M, Salanti G. In network meta-analysis, most of the information comes from indirect evidence: empirical study. *J Clin Epidemiol*. 2020;120:42-49.
117. 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;33:607-617.
118. Zhang J, Carlin BP, Neaton JD, et al. Network meta-analysis of randomized clinical trials: reporting the proper summaries. *Clin Trials*. 2014;11:246-262.
119. Hong H, Chu H, Zhang J, Carlin BP. A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Res Synth Methods*. 2016;7:6-22.
120. Piepho HP, Williams ER, Madden LV. The use of two-way linear mixed models in multitreatment meta-analysis. *Biometrics*. 2012;68:1269-1277.
121. White I, Turner R, Karahalios A, Salanti G. A comparison of arm-based and contrast-based models for network meta-analysis. *Stat Med*. 2019;38:5197-5213.
122. Senn S, Gavini DM, Scheen A. Issues in performing a network meta-analysis. *Stat Methods Med Res*. 2013;22:169-189.
123. Dias S, Ades AE. Absolute or relative effects? Arm-based synthesis of trial data. *Res Synth Methods*. 2016;7(1):23-28.
124. Dias S, Welton NJ, Sutton AJ, Ades AE. Evidence synthesis for decision making 5: the baseline natural history model. *Med Decis Making*. 2013;33:657-670.
125. Bucher HC, Griffith L, Guyatt GH, Opravil M. Meta-analysis of prophylactic treatments against pneumocystis Carinii pneumonia and toxoplasma encephalitis in HIV-infected patients. *J AIDS Human Retrovir*. 1997;15:104-114.
126. White I, Barrett J, Jackson D, Higgins J. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. *Res Synth Methods*. 2012;3:111-125.
127. 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;33:641-656.
128. Jackson D, Law M, Barrett JK, et al. Extending DerSimonian and Laird's methodology to perform network meta-analyses with random inconsistency effects. *Stat Med*. 2015;35:819-839.

129. Rhodes KM, Turner RM, Higgins JPT. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *J Clin Epidemiol*. 2015; 68:52-60.
130. Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JPT. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Stat Med*. 2015;34(6):984-998.
131. National Institute for Health and Care Excellence. Depression in Adults: Treatment and Management. NICE Guideline [NG 222]. London. 2022.
132. Brignardello-Petersen R, Mustafa RA, Reed AC, et al. GRADE approach to rate the certainty from a network meta-analysis: addressing incoherence. *J Clin Epidemiol*. 2019;109:77-85.
133. Hultcrantz M, Rind D, Akl EA, et al. The GRADE working group clarifies the construct of certainty of evidence. *J Clin Epidemiol*. 2017;87:4-13.
134. Brignardello-Petersen R, Bonner A, Alexander P, et al. Advances in the GRADE approach to rate the certainty in estimates from a network meta-analysis. *J Clin Epidemiol*. 2018; 93:36-44.
135. Brignardello-Petersen R, Guyatt GH, Mustafa RA, Chua DK, Hultcrantz M, Schünemann HJ. GRADE guidelines 33: addressing imprecision in a network meta-analysis. *J Clin Epidemiol*. 2012;139:49-56.
136. Salanti G, Del Giovane C, Chaimani A, Caldwell DM, Higgins JPT. Evaluating the quality of evidence from a network meta-analysis. *PloS One*. 2014;9(7):e99682.
137. Caldwell DM, Ades A, Dias S, et al. A threshold analysis assessed the credibility of conclusions from network meta-analysis. *J Clin Epidemiol*. 2016;80:68-76.
138. Phillippo DM, Dias S, Welton NJ, Caldwell DC, Taske N, Ades AE. Threshold analysis as an alternative to GRADE for assessing confidence in guideline recommendations based on network meta-analyses. *Ann Intern Med*. 2019;170:538-546.
139. Phillippo DM, Dias S, Ades AE, Didelez V, Welton NJ. Sensitivity of treatment recommendations to bias in network meta-analysis. *J R Stat Soc A Stat Soc*. 2018;181(3):843-867.
140. Savović J, Jones HE, Altman DG, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Ann Intern Med*. 2012;157: 429-438.
141. Elliott JH, Turner T, Clavisi O, et al. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS Med*. 2014;11(2):e1001603.
142. National Institute for Health and Care Excellence. Acne vulgaris: management. NICE Guideline [NG 198]. London. 2021.
143. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *Am J Epidemiol*. 2010;172:107-115.
144. Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. *J Royal Stat Soc Series A, (Stat Soc)*. 2011;174: 369-386.
145. Rosenbaum PR. Model-based direct adjustment. *J Stat Assoc Am*. 1987;82:387-395.
146. Zhang Z, Nie L, Soon G, Hu Z. New methods for treatment effect calibration, with applications to non-inferiority trials. *Biometrics*. 2016;72(1):20-29.
147. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. *NICE DSU Technical Support Document 18: Methods for Population-Adjusted Indirect Comparisons in Submissions to NICE*. ScHARR, University of Sheffield; 2016.
148. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for population-adjusted indirect comparisons in health technology appraisal. *Med Decis Making*. 2018; 38(2):200-211.
149. Varadhan R, Henderson NC, Weiss CO. Cross-design synthesis for extending the applicability of trial evidence when treatment effect is heterogeneous: part I. Methodology. *Commun Stat Case Studies Data Anal Appl*. 2017;2(3–4):112-126.
150. Signorovitch JE, Wu EQ, Yu AP, et al. Comparative effectiveness without head-to-head trials a method for matching-adjusted indirect comparisons applied to psoriasis treatment with Adalimumab or Etanercept. *Pharmacoeconomics*. 2010; 28(10):935-945.
151. Ishak JK. Indirect treatment comparison without network meta-analysis: overview of novel techniques. *Evidera*. 2014:1-6.
152. Caro JJ, Ishak KJ. No head-to-head trial? Simulate the missing arms. *Pharmacoeconomics*. 2010;28(10):957-967.
153. Phillippo DM, Dias S, Ades AE, Welton NJ. Assessing the performance of population adjustment methods for anchored indirect comparisons: a simulation study. *Stat Med*. 2020; 39(30):4885-4991.
154. Kahan BC, Jairath V, Doré CJ, Morris TP. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials*. 2014;15(139):1-7.
155. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. *Int J Epidemiol*. 1989;18(1):269-274.
156. Riley RD, Phillippo DM, Dias D. Network meta-analysis using IPD. In: Riley RD, Tierney JF, Stewart LA, eds. *Individual data meta-analysis*. Wiley; 2021.
157. Jackson C, Best N, Richardson S. Improving ecological inference using individual-level data. *Stat Med*. 2006;25:2136-2159.
158. Jackson C, Best N, Richardson S. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *J Royal Stat Soc Series A, (Stat Soc)*. 2008;171:159-178.
159. Jansen JP. Network meta-analysis of individual and aggregate level data. *Res Synth Methods*. 2012;3(2):177-190.
160. Phillippo DM, Ades AE, Belger M, et al. Multilevel network meta-regression for population-adjusted treatment comparisons. *J Royal Stat Soc Series A, (Stat Soc)*. 2020;183(3):1189-1210.
161. Donegan S, Williamson P, D'Alessandro U, Garner P, Tudor SC. 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.
162. Saramago P, Sutton AJ, Cooper NJ, Manca A. Mixed treatment comparisons using aggregate and individual participant level data. *Stat Med*. 2012;31:3516-3536.
163. Hamza T, Chalkou K, Pellegrini F, et al. Synthesizing cross-design evidence and cross-format data using network meta-regression. *Res Synth Methods*. 2023;14(2):283-300.
164. Phillippo DM, Dias S, Ades AE, et al. Validating the assumptions of population adjustment: application of multilevel network meta-regression to a network of treatments for plaque psoriasis. *Med Decis Making*. 2023;43(1):53-67.

165. Liu Y, Xia C. Fundamental equations of BLUE and BLUP in the multivariate linear model with applications. *Commun Stat Theory Methods*. 2013;42(3):398-412.
166. Jackson D. Alternative weighting schemes when performing matching-adjusted indirect comparisons. *Res Synth Methods*. 2021;12:333-346.
167. Remiro-Azócar A, Heath A, Baio G. Parametric G-computation with limited individual patient data. *Res Synth Methods*. 2022; 13(6):716-744.
168. Welton NJ, Phillippo DM, Owen R, et al. *NICE Decision Support Unit. CHTE2020 Sources and Synthesis of Evidence: Update to Evidence Synthesis Methods*. University of Sheffield; 2020.
169. Wolfson M, Wallace SE, Masca M, et al. DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *Int J Epidemiol*. 2010;39(5):1372-1382.
170. Gaye A, Marcon Y, Isaeva J, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. *Int J Epidemiol*. 2014;43(6):1929-1944.
171. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA*. 1995;273(5):408-412.
172. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *Br Med J*. 2008;336:601-605.
173. Savović J, Jones H, Altman D, et al. Influence of study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. *Health Technol Assess*. 2012;16(35):1-82.
174. Savović J, Turner RM, Mawdsley D, et al. Association between risk-of-bias assessments and results of randomized trials in Cochrane reviews: the ROBES meta-epidemiologic study. *Am J Epidemiol*. 2018;187(5):1113-1122.
175. Welton NJ, Ades AE, Carlin J, Altman DG, Sterne JAC. Models for potentially biased evidence in meta-analysis using empirically based priors. *J Royal Stat Soc Series A, (Stat Soc)*. 2009;172(1):119-136.
176. Dias S, Welton NJ, Marinho VCC, Salanti G, Higgins JPT, Ades AE. Estimation and adjustment of bias in randomised evidence by using mixed treatment comparison meta-analysis. *J Royal Stat Soc (A)*. 2010;173(3):613-629.
177. Salanti G, Dias S, Welton NJ, et al. Evaluating novel agent effects in multiple treatments meta-regression. *Stat Med*. 2010;29:2369-2383.
178. Moreno SG, Sutton AJ, Turner EH, et al. Novel methods to deal with publication biases: secondary analysis of antidepressant trials in the FDA trial registry database and related journal publications. *BMJ*. 2009;339:b2981.
179. Mavridis D, White IR, Higgins JPT, Cipriani A, Salanti G. Allowing for uncertainty due to missing continuous outcome data in pairwise and network meta-analysis. *Stat Med*. 2015;34:721-741.
180. White IR, Wood A, Welton NJ, Ades AE, Higgins JPT. Allowing for uncertainty due to missing data in meta-analysis: part 2: hierarchical models. *Stat Med*. 2008;27:728-745.
181. Schmitz S, Adams R, Walsh C. Incorporating data from various trial designs into a mixed treatment comparison model. *Stat Med*. 2013;32(17):2935-2949.
182. Efthimiou O, Mavridis D, Debray TPA, et al. Combining randomized and non-randomized evidence in network meta-analysis. *Stat Med*. 2017;36(8):1210-1226.
183. Zhang K, Arora P, Sati N, et al. Characteristics and methods of incorporating randomized and nonrandomized evidence in network meta-analyses: a scoping review. *J Clin Epidemiol*. 2019;113:1-10.
184. Hussein H, Abrams KR, Gray LJ, Anwer S, Dias S, Bujkiewicz S. Hierarchical network meta-analysis models for synthesis of evidence from randomised and non-randomised studies. *BMC Med Res Methodol*. 2023;23:97.
185. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database Syst Rev*. 2014;2014:4.
186. Verde PE, Ohmann C. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. *Res Synth Methods*. 2015;6(1):45-62.
187. Nikolaidis GF, Woods B, Palmer S, Soares MO. Classifying information-sharing methods. *BMC Med Res Methodol*. 2021; 21:13.
188. Ibrahim J, Chen M-H. Power prior distributions for regression models. *Stat Sci*. 2000;15(1):46-60.
189. Spiegelhalter DJ, Best NG. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Stat Med*. 2003;22(23):3687-3709.
190. Prevost TC, Abrams KR, Jones DR. Hierarchical models in generalised synthesis of evidence: an example based on studies of breast cancer screening. *Stat Med*. 2000;19:3359-3376.
191. Dias S, Welton NJ, Sutton AJ, Ades AE. Evidence synthesis for decision making 1: introduction. *Med Decis Making*. 2013; 33:597-606.
192. Verde PE. A bias-corrected meta-analysis model for combining, studies of different types and quality. *Biom J*. 2021;63: 406-422.
193. Lindley DV. Dynamic programming and decision theory. *Appl Stat*. 1961;10:39-51.
194. Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of heath care technologies. *J Health Econ*. 1999;18:341-364.
195. Brignardello-Petersen R, Florez ID, Izcovich A, et al. GRADE approach to drawing conclusions from a network meta-analysis using a minimally contextualised framework. *BMJ*. 2020;371:m3900.
196. Brignardello-Petersen R, Izcovich A, Rochwerg B, et al. GRADE approach to drawing conclusions from a network meta-analysis using a partially contextualised framework. *BMJ*. 2020;371:m3907.
197. Salanti G, Nikolakopoulou A, Efthimiou O, Mavridis D, Egger M, White IR. Introducing the treatment hierarchy question in network meta-analysis. *Am J Epidemiol*. 2021;191(5):930-938.
198. Papakonstantinou T, Salanti G, Mavridis D, Rucker G, Schwarzer G, Nikolakopoulou A. Answering complex hierarchy questions in network meta-analysis. *BMC Med Res Methodol*. 2022;22(47):47.
199. Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J Clin Epidemiol*. 2011;64:163-171.

200. Rücker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. *BMC Med Res Methodol*. 2015;15(58):1-9.
201. Mavridis D, Porcher R, Nikolakopoulou A, Salanti G, Ravaud P. Extensions of the probabilistic ranking metrics of competing treatments in network meta-analysis to reflect clinically important relative differences on many outcomes. *Biom J*. 2020;62:375-385.
202. Raiffa H. *Decision Analysis: Introductory Lectures on Choices under Uncertainty*. Addison-Wesley; 1961.
203. Pratt JW, Raiffa H, Schlaiffer R. *Introduction to Statistical Decision Theory*. Massachusetts Institute of Technology; 1995.
204. Claxton K. Bayesian approaches to the value of information: implications for the regulation of new pharmaceuticals. *Health Econ*. 1999;8:269-274.
205. Claxton K. OFT, VBP: QED? *Health Econ*. 2007;16:545-558.
206. Kirwin E, Paulden M, McCabe C, Round J, Sutton M, Meacock R. The risk-based price: incorporating uncertainty and risk attitudes in health technology pricing (June 16 2023). *Social Science Research Network*, 2023. https://ssrn.com/abstract=3956084 or 10.2139/ssrn.3956084
207. Griffin SC, Claxton KP, Palmer SJ, Sculpher MJ. Dangerous omissions: the consequences of ignoring uncertainty. *Health Econ*. 2011;20:212-224.
208. Lash TL, Fox MP, Cooney D, Lu Y, Forshee RA. Quantitative bias analysis in regulatory settings. *Am J Public Health*. 2016; 107(7):1227-1230.
209. Freeman SC, Sutton AJ, Cooper NJ. Uptake of methodological advances for synthesis of continuous and timeto-event outcomes would maximize use of the evidence base. *J Clin Epidemiol*. 2020;124:94-105.
210. Riley R. Multivariate meta-analyis: the effect of ignoring within-study correlation. *J Royal Stat Soc Series A (Stat Soc)*. 2009;172:789-811.
211. Mavridis D, Salanti G. A practical introduction to multivariate meta-analysis. *Stat Methods Med Res*. 2013;22(2):133-158.
212. Shadish WR, Haddock CK. Combining estimates of effect size. In: Cooper H, Hedges LV, eds. *The Handbook of Research Synthesis*. Russell Sage Foundation; 1994:261-281.
213. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med*. 1999;18:2693-2708.
214. 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; 33:618-640.
215. Caldwell DM, Welton NJ, Ades AE. Mixed treatment comparison analysis provides internally coherent treatment effect estimates based on overviews of reviews and can reveal inconsistency. *J Clin Epidemiol*. 2010;63:875-882.
216. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. *Stat Med*. 2010;29:932-944.
217. Rücker G. Network meta-analysis, electrical networks and graph theory. *Res Synth Methods*. 2012;3(4):312-324.
218. Ades AE, Caldwell DM, Reken S, Welton NJ, Sutton AJ, Dias S. Evidence synthesis for decision making 7: a reviewer's checklist. *Med Decis Making*. 2013;33:679-691.
219. 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.
220. Pedder H, Dias S, Bennetts M, Boucher M, Welton NJ. Joining the dots: linking disconnected networks of evidence using dose-response model-based network meta-analysis. *Med Decis Making*. 2021;41(2):194-208.
221. Petropoulou M, Efthimiou O, Rücker G, et al. A review of methods for addressing components of interventions in meta-analysis. *PloS One*. 2021;16(2):e0246631.
222. Welton NJ, Caldwell DM, Adamopoulos E, Vedhara K. Mixed treatment comparison meta-analysis of complex interventions: psychological interventions in coronary heart disease. *Am J Epidemiol*. 2009;169(9):1158-1165.
223. Freeman SC, Scott NW, Powell R, Johnston M, Sutton AJ, Cooper NJ. Component network meta-analysis identifies the most effective components of psychological preparation for adults undergoing surgery under general anesthesia. *J Clin Epidemiol*. 2018;98:105-116.
224. Mills E, Druyts E, Ghement I, Puhan MA. Pharmacotherapies for chronic obstructive pulmonary disease: a multiple treatment comparison meta-analysis. *Clin Epidemiol*. 2011;3:107-129.
225. Soares MO, Dumville J, Ades AE, Welton NJ. Treatment comparisons for decision making: facing the problems of sparse and few data. *J Royal Stat Soc Series A (Stat Soc)*. 2014;177: 259-279.
226. Owen RK, Tincello DG, Keith RA. Network meta-analysis: development of a three-level hierarchical modeling approach incorporating dose-related constraints. *Value Health*. 2015; 18(1):116-126.
227. Mayo-Wilson E, Dias S, Mavranezouli I, et al. Psychological and pharmacological interventions for social anxiety disorder in adults: a systematic review and network meta-analysis. *Lancet Psychiatry*. 2014;1:368-376.
228. Rhodes KM, Turner RM, Higgins JPT. Empirical evidence about inconsistency among studies in a pair-wise meta-analysis. *Res Synth Methods*. 2016;7(4):346-370.
229. Wei Y, Higgins J. Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. *Stat Med*. 2013;32:1191-1205.
230. Schmid CH, Trikalinos TA, Olkin I. Bayesian network meta-analysis for unordered categorical outcomes with incomplete data. *Res Synth Methods*. 2014;5:162-185.
231. Wei Y, Higgins JPT. Bayesian multivariate meta-analysis with multiple outcomes. *Stat Med*. 2013;32:2911-2934.
232. Efthimiou O, Mavridis D, Riley R, Cipriani A, Salanti G. Joint synthesis of multiple correlated outcomes in networks of interventions. *Biostatistics*. 2014;16(1):84-97.
233. Nam I-S, Mengerson K, Garthwaite P. Multivariate meta-analysis. *Stat Med*. 2003;22:2309-2333.
234. Jackson D, Riley R, White IR. Multivariate meta-analysis: potential and promise. *Stat Med*. 2011;30:2481-2598.
235. Bujkiewicz S, Thompson JR, Sutton AJ, et al. Use of Bayesian multivariate meta-analysis to estimate the HAQ for mapping onto the EQ-5D questionnaire in rheumatoid arthritis. *Value Health*. 2014;17:109-115.
236. Price MJ, Blake HA, Kenyon S, et al. Empirical comparison of univariate and multivariate meta-analyses in cochrane

pregnancy and childbirth reviews with multiple binary outcomes. *Res Synth Methods*. 2019;10:440-451.

237. Riley RD, Jackson D, Salanti G, et al. Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. *Br Med J*. 2017;258:j3932.
238. Guyatt GH, Thorlund K, Oxman AD, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles—continuous outcomes. *J Clin Epidemiol*. 2013;66(2):173-183.
239. Ades AE, Lu G, Dias S, Mayo-Wilson E, Kounali D. Simultaneous synthesis of treatment effects and mapping to a common scale: an alternative to standardisation. *Res Synth Methods*. 2015;6:96-107.
240. Friedrich JO, Adhikari KJ, Beyenegh J. Ratio of means for analyzing continuous outcomes in meta-analysis performed as well as mean difference methods. *J Clin Epidemiol*. 2011; 64(5):556-564.
241. Lu G, Kounali D, Ades AE. Simultaneous multi-outcome synthesis and mapping of treatment effects to a common scale. *Value Health*. 2014;17:280-287.
242. Lu G, Ades AE, Sutton AJ, Cooper NJ, Briggs AH, Caldwell DM. Meta-analysis of mixed treatment comparisons at multiple follow-up times. *Stat Med*. 2007;26(20):3681-3699.
243. Stettler C, Allemann S, Wandel S, et al. Drug eluting and bare metal stents in people with and without diabetes: collaborative network meta-analysis. *Br Med J*. 2008;337(7671): 1331.
244. Jansen JP, Vieira MC, Cope S. Network meta-analysis of longitudinal data using fractional polynomials. *Stat Med*. 2015; 34:2294-2311.
245. Kirkham JJ, Riley RD, Williamson PR. A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews. *Stat Med*. 2012;31:2179-2195.
246. Sargent DJ, Wieand HS, Haller DG, et al. Disease-free survival versus overall survival As a primary end point for adjuvant colon cancer studies: individual patient data from 20,898 patients on 18 randomized trials. *J Clin Oncol*. 2005;23(34):8664-8670.
247. Hughes MD. Practical issues arising in an exploratory analysis evaluating progression-free survival as a surrogate endpoint for overall survival in advanced colorectal cancer. *Stat Methods Med Res*. 2008;17(5):487-495.
248. Flaherty KT, Hennig M, Lee SJ, et al. Surrogate endpoints for overall survival in metastatic melanoma: a meta-analysis of randomised controlled trials. *Lancet Oncol*. 2014;15:297-304.
249. Daly CH, Maconachie R, Ades AE, Welton NJ. A non-parametric approach for jointly combining evidence on progression free and overall survival time in network meta-analysis. *Res Synth Methods*. 2022;13(5):573-584.
250. Guyot P, Ades AE, Ouwens MJNM, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2012;12:9.
251. Guyot P, Ades AE, Beasley M, Lueza B, Pignon J-P, Welton NJ. Extrapolation of survival curves from cancer trials using external information. *Med Decis Making*. 2016;37:353-366.
252. Latimer N. *NICE DSU Technical Support Document 14: Undertaking Survival Analysis for Economic Evaluations alongside Clinical Trials: Extrapolation with Patient-Level Data*. NICE Decision Support Unit; 2011.
253. Welton NJ, Willis SR, Ades AE. Synthesis of survival and disease progression outcomes for health technology assessment of cancer therapies. *Res Synth Methods*. 2010;1:239-257.
254. Welton NJ, Cooper NJ, Ades AE, Lu G, Sutton AJ. Mixed treatment comparison with multiple outcomes reported inconsistently across trials: evaluation of antivirals for treatment of influenza A and B. *Stat Med*. 2008;27:5620-5639.
255. Anwer S, Ades AE, Dias S. Joint synthesis of conditionally related multiple outcomes makes better use of data than separate meta-analyses. *Res Synth Methods*. 2020;11(4):496-506.
256. Cope S, Jansen JP. Quantitative summaries of treatment effect estimates obtained with network meta-analysis of survival curves to inform decision-making. *BMC Med Res Methodol*. 2013;13:147.
257. Freeman SC, Cooper NJ, Sutton AJ, Crowther MJ, Carpenter JR, Hawkins N. Challenges of modelling approaches for network meta-analysis of time-to-event outcomes in the presence of non-proportional hazards to aid decision making: application to a melanoma network. *Stat Methods Med Res*. 2022;31(5):839-861.
258. Rutherford MJ, Lambert PC, Sweeting MJ, et al. NICE DSU Technical Support Document 21. Flexible methods for survival analysis. 2020. http://www.nicedsu.org.uk
259. Ouwens MJNM, Philips Z, Jansen JP. Network meta-analysis of parametric survival curves. *Res Synth Methods*. 2010;1:258-271.
260. Jansen JP. Network meta-analysis of survival data with fractional polynomials. *BMC Med Res Methodol*. 2011;11(1):1-14.
261. Wei Y, Royston P, Tierney JF, Parmar MKB. Meta-analysis of time-to-event outcomes from randomized trials using restricted mean survival time: application to individual participant data. *Stat Med*. 2015;34:2881-2898.
262. Freeman SC, Carpenter JR. Bayesian one-step IPD network meta-analysis of time-to-event data using Royston-Parmar models. *Res Synth Methods*. 2017;8:451-464.

AUTHOR BIOGRAPHIES

A. E. Ades is Professor Emeritus at the University of Bristol, working on evidence synthesis for decision making and epidemiology. He received the Ingram Olkin award for Lifetime Achievement in Research Synthesis Methodology in 2010.

Nicky J. Welton is Professor of Statistical and Health Economic Modelling at the University of Bristol working on methods for evidence synthesis in healthcare decision-making. She is co-Director of the Guidelines Technical Support Unit for NICE, and Co-Director of the Bristol Technology Assessment Group.

Sofia Dias is Professor in Health Technology Assessment at the University of York working on methods for evidence synthesis in healthcare decision-making. She is Director of the NIHR funded Centre for Reviews and Dissemination/Centre for Health Economics Technology Assessment Reviews Group.

David M. Phillippo is Research Fellow in Evidence Synthesis at the University of Bristol. His research focuses on methods for network meta-analysis and population adjustment.

Deborah M. Caldwell is Professor of Epidemiology and Public Health at the University of Bristol, working on evidence synthesis for decision making and epidemiology. She is Co-Director of the NIHR funded Bristol Evidence Synthesis Group and Co-Convenor of the Cochrane Comparing Multiple Interventions Methods Group.

SUPPORTING INFORMATION

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

How to cite this article: Ades AE, Welton NJ, Dias S, Phillippo DM, Caldwell DM. Twenty years of network meta-analysis: Continuing controversies and recent developments. *Res Syn Meth*. 2024;1-26. doi:10.1002/jrsm.1700.