CHTE2020 SOURCES AND SYNTHESIS OF EVIDENCE; UPDATE TO EVIDENCE SYNTHESIS METHODS

REPORT BY THE DECISION SUPPORT UNIT

20 April 2020

Welton NJ[1], Phillippo DM[1], Owen R[2], Jones HE[1], Dias S[3], Bujkiewicz S[2], Ades AE, Abrams KR[2]

1. University of Bristol
2. University of Leicester
3. University of York

Decision Support Unit, ScHARR, University of Sheffield, Regent Court, 30 Regent Street
Sheffield, S1 4DA

Tel (+44) (0)114 222 0734
E-mail dsuadmin@sheffield.ac.uk
Website www.nicedsu.org.uk
Twitter @NICE_DSU

ABOUT THE DECISION SUPPORT UNIT

The Decision Support Unit (DSU) External Assessment Centre is based at the University of Sheffield with members at York, Bristol, Leicester and the London School of Hygiene and Tropical Medicine. The DSU is commissioned by The National Institute for Health and Care Excellence (NICE) to provide a research and training resource to support the Institute's Centre for Health Technology Evaluation Programmes. Please see our website for further information www.nicedsu.org.uk.

The production of this document was funded by the National Institute for Health and Care Excellence (NICE) through its Decision Support Unit. The views, and any errors or omissions, expressed in this document are of the authors only. NICE may take account of part or all of this document if it considers it appropriate, but it is not bound to do so.

Acknowledgements

Dr Nicholas Latimer, University of Sheffield and Dr Edit Remak, Visible Analytics Limited for useful discussions regarding treatment switching. Sumayya Anwer (University of York) helped identify references.

Author Contributions

Authors are listed in reverse alphabetical order. NJW led sections 2.4, 6.1 - 6.4, and 6.6, contributed to sections 2.1-2.5 and 5.1, compiled and reviewed the full report, and managed the project overall. DMP contributed to sections 2.1-2.3 and 7. RO contributed to section 8. HEJ led section 8. SD led sections 2.5, 2.6, 3.2, and 5.1, and contributed to sections 2.1-2.4, and 7. SB led sections 3.1, 4, and 5.2. AEA led sections 2.1-2.3 and 7, contributed to section 8, and shaped the project in its early stages. KRA led section 6.5.

This report should be referenced as follows:

Welton, N.J., Phillippo, D.M., Owen, R., Jones, H.J., Dias, S., Bujkiewicz, S., Ades, A.E., Abrams, K.R. DSU Report. CHTE2020 Sources and Synthesis of Evidence; Update to Evidence Synthesis Methods. March 2020

EXECUTIVE SUMMARY

In this document we critically review existing and emerging methods for synthesising evidence on clinical effectiveness for decision-making in Health Technology Appraisals (HTA). We focus on methods that have been developed or have been applied in HTA since the 2013 NICE Guide to the Methods of Technology Appraisal. The following methodological areas are reviewed:

- network meta-analysis, including a comparison of different parameterisations, modelling treatment effects, bias-adjustment, and methods to assess reliability of the recommendations based on network meta-analysis
- adjustment for population differences in indirect treatment comparisons and network meta-analysis, distinguishing between the case where networks are connected or disconnected
- synthesis of observational and randomised evidence
- multiple outcomes, including multivariate meta-analysis models and models that exploit structural relationships in the evidence
- synthesis of evidence on surrogate outcomes for clinical endpoints
- use of informative priors for between study heterogeneity and correlation parameters
- synthesis of data on survival including the cases when proportional hazards does or does not hold, joint modelling of progression free survival and overall survival, adjustment for patients switching treatments on disease progression, and synthesis of RCT and external data to aid extrapolation
- synthesis of evidence on the comparative accuracy of diagnostic tests, including the cases where there is a single test or multiples tests, and the situation where there is no gold standard.

We make recommendations for updates to the NICE Guide to the Methods of Technology Appraisal for each topic, as well as recommendations for further research and future technical support documents.

1 CONTENTS

2 NETWORK META-ANALYSIS (NMA)

2.1 Aggregate NMA with alternative parameterisations

In this section we use terminology from the statistical literature which sometimes conflicts with terminology used in the meta-analysis literature. We use *fixed effects for trials* to refers to separate (unconstrained) baseline effects for each study. *Common treatment effect* refers to the same relative treatment effect across studies (which is what is usually termed a *fixed effect meta-analysis* model in the meta-analysis literature). Furthermore, we use the terminology *arm-based models* to refer to models which put the NMA model on treatment arms, and *contrast based models* to refer to models which put the NMA model on relative effects. Note that this terminology is distinct from the terminology used to describe the data used to inform the models. Contrast-based models can be estimated using arm-level or contrast-level data. Arm-based models can only be estimated using arm-level data.

The NMA models usually appearing in submissions to NICE have been based on models described in TSD-2 (Dias et al., 2013, Lu and Ades, 2006) which also provides WinBUGS Coding and further elaboration. These models are *contrast-based* (CB) with *hierarchical models for relative treatment effects*, and *fixed (unrelated) effects for trial baseline effects.*

Several alternative parameterisations of NMA models for aggregate data have been proposed, summarised in a recent review (White et al., 2019). Here we look at four types of variants:

- Hierarchical models for trial effects
- Arm-based models
- Models for all treatment arms, not just those observed
- Inconsistency models

2.1.1 Hierarchical models for trial baseline effects

Key references: (White et al., 2019, Senn et al., 2013)

In the standard NMA models (Dias et al., 2018) separate fixed effects are given for each trial baseline on which relative effects are then added. Rather than fixed effects for trials, an alternative is to put a hierarchical model on trial effects (also referred to as trial baselines or intercepts) in addition to the hierarchical model on relative treatment effects. This has been said to *compromise randomisation (White et al., 2019)*, meaning that it may introduce bias into the relative treatment effects. Whilst it might be expected that in practice the estimated relative treatment effects might not greatly affected (Senn et al., 2013), it has been demonstrated that quite large biases could occur in certain circumstances (White et al., 2019). Because fixed effect baselines guard against the possibility of such biases, and code is readily available (Evidence Synthesis TSDs), we believe that this remains the most prudent approach. This is also in line with standard practise for pair-wise meta-analyses. The fixed trial effect NMA models of TSD-2 are generalisations of standard pairwise meta-analysis: both produce exactly the same results, as long as between-arm correlations in the likelihood for multi-arm trials are handled correctly (Franchini et al., 2012).

There are circumstances with sparse networks and trials with zero cells the TSD-2 models will not converge. If this is still an issue after removing studies with zero events on all arms (which provide no information on relative effects) and treatments where the event is not possible ((Dias et al., 2018) Chapter 6), further modelling options to obtain stable computation are available (Dias et al. 2018 Ch6). Options include placing a random effects model on the trial baseline effects (Ohlssen et al., 2014), use of informative priors on treatment effect parameters which may come from observational data (Soares et al., 2014), or borrowing strength through a model on treatment effects (see section 1.4). Each of these methods makes assumptions, borrowing strength from exchangeability of trial baselines, external sources, and treatment models respectively. A critique of those different assumptions is required in each application to guide as to the most relevant approach. Another option is to consider the scale on which the model is applied, for example using risk-differences, or rate-ratios rather than odds-ratios for the relative effects.

2.1.2 *Arm-based models*

Key references: (Dias and Ades, 2016, Hong et al., 2016a, Hong et al., 2016b, White et al., 2019, Zhang et al., 2014)

CB models have parameters representing the trial-specific relative treatment effects. In a*rm-based* (AB) models there are hierarchical models for the absolute effects on each arm, and correlation parameters between absolute effects. White et al. (2019) show that for every AB model a CB equivalent can be constructed with a specific correlation structure between the baselines and the treatment effects. However, the assumption of a hierarchical model on baselines already makes it difficult to justify the use of AB models in HTA, although there are several other reasons why they should be avoided.

First, the accepted practice with decision models for health interventions based on randomised trial evidence is that the baseline model is informed by absolute estimates of outcomes on a reference treatment. Standard practice in NICE submissions, and widely accepted in HTA, is that the identification of evidence sources to inform the baseline model is an entirely different process to identifying trials to inform relative treatment effects. Arm-based models, however, at least as they stand now, oblige the user to use the same set of trials to inform both the baseline model and the relative treatment effects (Dias and Ades, 2016).

A second difficulty is that, without relative effect parameters, the concept of "consistency" which applies to relative effects, and meta-regression, in which relative effects are regressed against covariates, become much more difficult to operationalise. This is important because the validity of estimates from an NMA rely on consistency of effects between direct and indirect evidence, and so it is essential to be able to assess this assumption. Inconsistency can be assessed in AB models using a form of meta-regression (White et al., 2019). However, standard solutions such as those in TSD-3 and TSD-4 would not be applicable, and AB versions of these analyses appear to have no advantages.

2.1.3 <u>Models for all treatments or just those included in a study?</u>

Key references: (Phillippo, 2019, White et al., 2019)

The NMA models in TSD-2 include parameters only for the treatment arms that have been included in a given study. For example, for a trial comparing treatments labelled as "2" and "3", there is a parameter for the relative effect of treatment 3 compared to treatment 2, but no parameters for the effects of 2 or 3 relative to the "missing" reference treatment "1" (Dias et al., 2013, Lu and Ades, 2006). These models are sometimes referred to as "baseline shift". An alternative formulation includes a parameter for the relative effects of treatments 2 and 3 compared to treatment 1, even though treatment 1 hasn't been included in that study (White et al. 2019), but the model is still estimated from arm-level data. The two models have the same basic parameters and are exactly equivalent ***provided that*** correlations between random effects in multi-arm trials are handled appropriately (Franchini et al., 2012). Both formulations can be used in submissions as long as the correlations are handled appropriately in the latter. The parameterization that includes all the treatment effects can be more convenient in certain applications including meta-regression, population-adjustment, and heterogeneous between-trial variances (Lu and Ades, 2009), but may also be slower to compute for random effects models due to the extra parameters.

A frequentist model for NMA implemented in Stata (White, 2015) "augments" the data to create a treatment "1" arm with very low precision in all studies that do not include treatment "1". This allows a baseline shift model to be estimated within a frequentist paradigm, and gives very similar results to the Bayesian implementation of the baseline shift model in the Evidence Synthesis TSDs unless networks are sparse (Dias et al. 2018).

2.1.4 Use of inconsistency models for decision making

Key references: (Dias et al., 2010b, Dias et al., 2011c, Higgins et al., 2012, Jackson et al., 2014, Lu and Ades, 2006)

A key assumption of NMA is consistency of relative treatment effects. The "null hypothesis" of consistency should always be tested wherever there are "loops" of evidence (Dias et al., 2011). Inconsistency models (Dias et al., 2010; Higgins et al., 2012; Jackson et al., 2014) abandon the "consistency assumption" and introduce additional inconsistency terms or treatment effects. Inconsistency models can be used to test the null hypothesis of consistency, by providing a global test for inconsistency by comparing goodness of fit between the inconsistency and consistency models (Dias et al., 2010). Inconsistency can be further explored by plotting the contribution to the posterior residual deviance of each data-point for the consistency versus the inconsistency model (the deviance-deviance plots (Dias et al., 2018)), where datapoints that are much more deviant under the consistency model indicating potential inconsistency/outliers.

A further proposal has been to use inconsistency models for decision making (Jackson et al., 2014), where inconsistency parameters are incorporated in much the same way that heterogeneity is in random effects models, by assuming that the inconsistency terms are exchangeable. Our view is that inconsistency models are not appropriate for decision making, firstly because the exchangeable inconsistency terms have no clear interpretation; secondly because it is difficult to imagine a mechanism that would generate them; and finally because taking at face value a body of data that is recognised as inconsistent may lead to perverse recommendations that are unlikely to gain support from all stakeholders. Where inconsistency is identified, then the first response should be to try to understand what might be causing it (data extraction errors, methodological differences between studies, lumping treatment definitions, etc) and adjust for it. If that is not possible then sensitivity analyses should be conducted to understand the potential impact of inconsistency on recommendations.

Note that NICE TAs often use indirect comparisons where there are no loops of evidence and it is not possible to test for inconsistency. However, in connected networks with loops of evidence it is essential to test for inconsistency.

2.1.5 Recommendations

- The Methods guide should be updated with a new section describing alternative parameterisations. This should clearly state that the current methods in TSD-2 are the preferred methods, and set out under what circumstances alternative methods may be justified.
- The methods for NMA in the Evidence Synthesis TSDs are recommended
- Hierarchical models on trial effects should only be used if no other alternative solutions are available, and the potential for bias should be clearly pointed out.
- Models which have parameters for all treatment contrasts that could have been included in a study can be used as long as the correlations are accounted for appropriately.
- Inconsistency should be checked if loops of evidence are present, and if inconsistency is identified an attempt should be made to explain and adjust for it. In the absence of an explanation, sensitivity analyses should be conducted. Inconsistency models should only be used for checking for inconsistency and not for decision making.

2.1.6 Research Recommendations

- A manuscript giving a practical guide for inconsistency checking is in preparation (early draft). 1 month's work to complete (once time found to work on it).

2.2 Population-adjustment methods in connected networks (anchored case)

Standard NMA assumes that the distribution of effect modifiers is the same in all trials. Population adjustment methods (Phillippo et al., 2016, Phillippo et al., 2018a) open up the possibility of relaxing this assumption, and thereby making a valid BvC comparison based on AvB and AvC trials with different distributions of effect modifiers. These methods (Signorovitch et al., 2010, Caro and Ishak, 2010) were developed for specific situation that arises in NICE TAs, where a manufacturer has access to IPD from its own trial (say the AvB trial), while only aggregate summary level data is available from the competitor's AvC trials, but a comparison between BvC is required.

2.2.1 Matching-adjusted Indirect comparisons (MAIC) – for connected networks ("anchored" comparisons)

Key references: (Phillippo et al., 2016, Signorovitch et al., 2012, Signorovitch et al., 2010)

Matched Adjusted Indirect Comparisons (MAIC) is a propensity score method proposed by Signorovitch et al. (2010). The method identifies covariates measured in the trial with IPD that interact with the treatment effect (for example using regression techniques). The selected covariates are then used to reweight the observations in the IPD study, using inverse propensity score weightings, to match the marginal covariate distribution in a study with aggregate data only. The weights are obtained using a method of moments approach because only marginal summaries of covariates in the aggregate data study are available (eg mean and standard deviation from a baseline characteristics table). A critique can be found in TSD-18. MAICs assume that all effect modifiers have been adjusted for, but this is limited by the covariates that have been measured and reported in both studies. Adjusting for covariates that are not effect modifiers reduces the precision of the estimates with no gain in reliability of the estimates. There is usually little evidence of effect modification within a single study, due to low power to detect effect modifiers. MAIC is only able to estimate treatment effects in the AvC (competitor's) trial population. Therefore, if the manufacturers of treatments B and C were to each submit their own MAICs, based on the same two trials, the results they would produce would be relevant for two different populations (which we suppose to have substantial differences as this is the motivation to conduct population adjustment in the first place). Furthermore, the target population for the decision may be a different population again. This could lead to different and potentially spurious recommendations. Finally, MAICs are problematic when there are more than 2 studies, because population adjustment is different for each study that is being adjusted to. This makes application of MAIC to networks of evidence impossible.

Furthermore, MAICs perform poorly in simulation studies, and in some scenarios perform worse than standard NMA with no population adjustment (Phillippo, 2019).

2.2.2 *<u>Simulated Treatment Comparisons (STCs) - for connected networks ("anchored" comparisons)</u>*

Key references: (Caro and Ishak, 2010, Ishak, 2014, Phillippo, 2019, Phillippo et al., 2016)

Simulated Treatment Comparisons (STC) is a regression-based covariate adjustment method. An outcome regression model is fitted to the covariates using IPD. This is then used to predict the outcome of treatment B in the AvC trial, but rather than plugging in the mean covariate values from the AvC trial, covariate values are simulated from a distribution with the reported marginal summaries, and the resulting predicted effect of treatment B calculated. These simulations are averaged over to generate a summary outcome for treatment B in the AvC trial. Further details can be found in the key references, and a critique in TSD-18 (Phillippo et al., 2016). The same criticisms of MAIC apply to STC (see section 1.2.1), but with the exception that STC performs better in simulation studies than MAIC for the 2 study scenario (Phillippo, 2019).

2.2.3 Multi-Level Network Meta-regression (ML-NMR) – for connected networks ("anchored" comparisons)

Key references: (Phillippo et al., 2020, Phillippo, 2019)

Recently a new approach, Multi-Level Network Meta-Regression (ML-NMR), has been developed (Phillippo, 2019; Phillippo et al., 2020). ML-NMR estimates relative treatment effects at two levels, the population level (which is the target for inference) and the IPD level, but it recognises the mathematical relation between the two, by integrating over the trial population. Integrating over the covariate distributions gives estimates that are more precise than those obtained by STC due to the additional uncertainty introduced by simulation.

ML-NMR has the advantage that it is applicable to any connected network with any mixture of IPD and aggregate data. It has the property that, in the absence of effect modifiers, it will produce the same results as standard NMA. Also, if all trials have IPD it gives the same result as IPD meta-regression, which is the gold standard method to population adjustment. This is a key property as it indicates that ML-NMR is underpinned at both ends by methods that are well-understood and which are readily replicable. This is not necessarily the case for MAICs which can be performed in a variety of different ways.

Like MAICs and STCs, ML-NMR also assumes that all effect modifiers have been adjusted for. The disadvantage of ML-NMR is that except for linear models with identity link and continuous covariates the integration step requires numerical methods. An R package calling Stan is currently being developed which includes models with log, cloglog, logit, and probit link functions, as well as models with survival analysis outcomes.

A recent simulation study (Phillippo, 2019) shows that ML-NMR performs similarly to STC in the 2-study scenario when the target population of interest is the population in the trial with aggregate data. However, it performs better than STC when the target population differs from that of the trial with aggregate data.

2.2.4 Recommendations

- TSD-18 (Phillippo et al., 2016) advises on circumstances under which MAIC and STC could be used in submissions, and sets out some particulars of how they should be used and presented.
- We recommend that a new TSD is prepared to show how to use ML-NMR, along with worked examples and software code, and that the Methods Guide is revised to make it clear that MAICs should not be used under any circumstances, that STCs can be can be used for two-study scenarios, and that ML-NMR is the preferred approach for anchored comparisons. This could be developed over the next 6 months.

2.3 Population-adjustment methods in disconnected networks (one-arm studies, the unanchored case)

2.3.1 <u>MAICs, STC, ML-NMR</u>

Key references: (Caro and Ishak, 2010, Ishak, 2014, Phillippo, 2019, Phillippo et al., 2016, Signorovitch et al., 2012, Signorovitch et al., 2010)

All of the population adjustment methods of section 1.2, can be applied to disconnected networks where there are two treatments to be compared but no common comparator. This can include the case where one of the studies is a single-armed trial. The ML-NMR method can also be applied to more general disconnected networks with multiple studies and treatments (Phillippo, 2019).

All of the general critiques of these methods (section 1.2) also apply to the unanchored case, but there is a far more important difficulty. With anchored comparisons the all population adjustment methods require that information is available on all effect modifiers, so that they can all be adjusted for. For unanchored comparisons a far stronger assumption required, that information is additionally available on *all the prognostic factors that are associated with the absolute outcome,* so that they can all be adjusted for. In practise it is unlikely that information on all prognostic factors and effect modifiers are available, and it is impossible to verify this. However, unanchored STC or ML-NMR may reduce the variation between absolute effects observed on the same treatment in different trials, whereas MAICs may in some circumstances actually increase variation.

The issue is therefore: how much reduction in variance can be achieved? More specifically, given a particular body of data in a submission, and perhaps taking into account other data in the literature, and/or other IPD held by a manufacturer from other trials, what methods can be applied to assess the likely degree of error in predictions regarding relative treatment effects based on unanchored comparisons.

TSD-18 (Phillippo et al., 2016) suggested some steps that should be taken in submissions to quantify the likely level of systematic error that occurs when unanchored population-adjusted treatment comparisons are made. These include out-of-sample methods, where multiple external evidence sources on one treatment A are available, and the between study heterogeneity of the observed estimates and the predicted population adjusted estimates are compared to obtain an estimate of the proportion of variability explained by the covariates. In-sample methods, such as, cross validation and other techniques from the causal inference literature are another option. Further work, funded by the MRC, is currently underway to develop these methods further.

2.3.2 Recommendations

- Unanchored indirect comparisons require all prognostic factors and effect modifiers to be adjusted for, which is unlikely to be achieved in practise. Whilst STC or ML-NMR can be used to reduce bias in unanchored comparisons, the potential for bias cannot be eliminated without further comparative, randomised research
- Whilst STC or ML-NMR are analytical tools that may be used to support decision making, the limitations of the use of these methods in this context need to be recognised, and if possible any systematic bias estimated using either out-of-sample or in-sample methods.

2.3.3 Research recommendation

- Research to identify the best approach to quantify the extent of error likely in submissions based on population-adjusted methods is needed, specifically in the unanchored case. A research project is underway to address this and will conducted over the next 24 months.

2.4 Models on treatment effects

How treatments are defined and modelled can have implications for the connectivity of a network, and also the precision of the estimates and potential for heterogeneity and inconsistency. Defining treatments at a high level of detail (eg dose, administration route, schedule, co-treatments) may mean that networks do not connect and analysis is not possible, or if networks do connect estimates may be very imprecise. Lumping treatments together that differ in key features that may interact with relative treatment effects may increase precision of estimates and connect networks, but at the cost of introducing heterogeneity and inconsistency so that the resulting estimates may be unreliable. An alternative is to use modelling to capture the structural differences between treatments, such as dose response models, hierarchical models for treatments within class, and component models when treatments can be defined as being a combination of constituent parts. Note that such models are distinct from standard meta-regression models because the treatment features that are being modelled represent different randomised groups, whereas in standard meta-regression covariates effects do not represent a randomised comparison (Dias et al., 2018)(Ch 8, section 8.6).

2.4.1 Dose models

Key references: (Mawdsley et al., 2016, Owen et al., 2015)

In NICE TAs interest is on comparisons between licensed doses only, however the evidence available may include comparisons with treatments at unlicensed doses and networks may be disconnected if this evidence is excluded. Two main approaches to modelling dose effects have been proposed: hierarchical models (Del Giovani et al., 2013, Owen et al., 2015) and dose-response models (Thorlund et al., 2014, Mawdsley et al., 2016).

Del Giovane et al. (2013) proposed a hierarchical model that assumes dose effects of the same agent are “similar” (exchangeable) but additionally allows a relationship with dose by modelling adjacent doses with a random walk process. However, the interpretation of estimates from these models can be difficult, in particular the model may give dose-response relationships that are not monotonically increasing or decreasing with dose, which is not plausible. Owen et al. (2015) overcome this limitation by introducing constraints on the model so that there is a monotonic relationship between different doses of the same agent. The Owen et al. (2015) approach has the advantage that the resulting estimates have face-validity (dose-response is monotonic with dose) and comparisons with specific doses that have been included in the trials can be made without making strong assumptions. The method does however assume that the dose effects are exchangeable across the doses that have been included in the trials, which may not be reasonable, especially if there is a large range of doses included which may have very different responses (this would be seen as high heterogeneity between doses within agent). Also, if there are not many different doses of an agent in the trials, then there is little to be gained by using a hierarchical model because there is little evidence with which to estimate between dose variability.

If there is sufficient evidence on a range of different doses for a particular agent, then functional relationships for dose-response models can be estimated. Mawdsley et al. (2016) provide a framework (Model-Based Network Meta-Analysis (MBNMA)) to fit a range of dose-response models including the Emax model which is commonly used in pharmacometric modelling. The MBNMA framework respects randomisation (unlike much of the Model-Based Meta-Analysis literature) (Mawdsley et al. 2016) and allows interpolation to predict responses for doses not included in the trials on which the model is estimated. An R package is available to fit these models (https://CRAN.R-project.org/package=MBNMAdose). The advantage of this approach is that it can connect otherwise disconnected networks and strengthen estimates (increased precision) between licensed doses of interest (Mawdsley et al. 2016). It is expected that more precise estimates would be obtained using MBNMA than the hierarchical approach of Owens et al. (2015), due to the parametric nature of the model. One limitation of MBNMA is that it requires evidence at a range of different doses for each agent where dose-response modelling is required. For use in NICE TAs this will mean that literature searches will need to be expanded to include phase-II evidence and head-to-head comparisons at non-licensed doses. Another limitation is that the method relies on an appropriate functional form for the dose-response curve to have been fitted, however a range of different functions can be estimated and model fit compared. Furthermore dose-response functions may be well understood based on known pharmacology of the agents and phase-II dose-response evidence which can be incorporated in the estimation and choice of model. The assumptions required for dose-response modelling may be less strong than those required for population adjustment for disconnected networks (section 1.3) so long as there is sufficient dose-response evidence available.

We are not aware that dose-models have been applied in NICE TAs to date.

2.4.2 Class models

Key references: Dias et al. (2018) Chapter 8, section 8.6.2, Owen et al (2015)

Class models can be used when the treatments in a network fall into a set of treatment classes with similar mechanisms of action, and have been applied in a range of clinical areas (Haas et al., 2012, Kew et al., 2014, Mayo-Wilson et al., 2014, Dakin et al., 2011, Warren et al., 2014, Soares et al., 2014, Dominici et al., 1999). Class models assume a hierarchical structure with treatments nested with class, so that an overall class-level mean effect and between treatment within class variance are estimated. These models have the property that if the between treatment within class variance is zero, then this is equivalent to lumping all treatments in the same class together, whereas a very high between treatment within class variance is equivalent to treating each treatment separately. The class model usually represents a compromise between these extremes where the between treatment within class variance reflects the degree of similarity and hence borrowing of strength of effects within a class. Class models are useful when data are sparse and NMA is otherwise impossible (disconnected networks) or gives imprecise estimates. Class models make sense when the mechanism of action and treatment effects are expected to be similar across treatment in the same class. A limitation is that the effects of treatments within a class are "shrunken" towards the class mean, which may disadvantage those treatments that are most effective within their class (or conversely enhance effect estimates for those that are least effective within their class). Another limitation is that there is often insufficient evidence with which to estimate between treatment within class variance, and this variation may be indistinguishable from the between study variance in random effect models (lack of identifiability). Ideally, each class would have its own between treatment variance parameter, but in practise there is often insufficient evidence to estimate these and they are shared across classes. Informative priors for the between treatment within class standard deviation could be used in the same way as priors for between study standard deviation (see section 4.1), but we are not aware of any evidence-based priors having been developed. Class models estimate both treatment and class level effects. A question therefore arises as to what the most appropriate summary effect is to use in an economic model. There are three options. If the decision needs to be made for each treatment within a class, then shrunken treatment-level estimates are appropriate. If the decision is for a typical treatment within the class, then the class mean estimate is appropriate. If the decision is for a random treatment within the class, then the predictive distribution is the most appropriate summary.

Class models have been used in NICE TAs previously, for example TA383 TNF-alpha inhibitors for ankylosing spondylitis and non-radiographic axial spondyloarthritis (https://www.nice.org.uk/guidance/ta383) (Corbett et al., 2016), where the predictive distribution was used to summarise the class effect in the economic model. Another example is TA527 beta interferons and glatiramer acetate for treating multiple sclerosis (https://www.nice.org.uk/Guidance/TA527), where a pooled model of effects from the risk sharing scheme data was accepted by the committee, and also used this to enable a comparison with Extavia which was not included in the risk sharing scheme. It is important to be aware that manufacturers may feel their treatment is disadvantages if the effects are shrunken towards a class mean effect, and discussion of the clinical validity of class effects is essential.

There have also been TAs where it is assumed that the relative effectiveness of treatments in the same class are equal (a fixed class model with zero between treatment variability within class). For example TAs in renal cell carcinoma have assumed that the relative effects for sunitinib and pazopanib are equal (eg TA542 https://www.nice.org.uk/guidance/ta542). If more treatments in this class were available in the future, class effect models could be used as an alternative. As mentioned above in TA527 Extavia was assumed to have the same effect as Betainterferon.

Class models can also be combined with models for dose (section 1.4.1). Owen et al (2015) developed a 3-level hierarchical model which modelled doses within agent, and agents within class. Mawdsley et al (2016) explored class models for the parameters of dose-response functions in MBNMA, and this modelling option is available in their R package (https://CRAN.R-project.org/package=MBNMAdose).

2.4.3 Component models

Key references: (Welton et al., 2009b, Mills et al., 2012)

Some interventions can be considered to be a sum of components parts (Melendez-Torre et al., 2015). For example psychological interventions are often considered to consist of components such as behavioural, cognitive, psycho-education, relaxation etc. (Welton et al., 2009b). More relevant for NICE TAs are combinations of drug treatments with different modes of action, for example cytotoxic chemotherapy plus immunotherapeutic therapies in oncology. Standard NMA methods can be applied where each combination of components is treated as if it were a distinct treatment, and this would be the usual approach in NICE TAs. However, when networks do not connect or estimates are very imprecise, then it may be reasonable to assume a model for intervention components.

(Welton et al., 2009b) proposed a series of network meta-regression models. The simplest is an additive model which assumes that for each intervention the effect of the combination is the sum of the effects of the component parts, so that there are no synergistic or antagonistic (interaction) effects. This model can be extended by adding two-way interactions, then 3-way interactions etc. The full interaction model is equivalent to standard NMA with each combination of components as a distinct intervention. In practise it is rarely possible to estimate interaction effects, so additive models are most commonly assumed. The method has been applied in TA244 (https://www.nice.org.uk/guidance/ta244) (National Institute for

Health and Care Excellence, 2012, Riemsma et al., 2011, Mills et al., 2012) where the effect of Long-active β2-agonists (LABA) plus Inhaled corticosteroids (ICS) were assumed to be the sum of the LABA and ICS effects. Recently, a frequentist implementation of the Welton et al. 2009 models has been developed in R (Rucker et al., 2019).

The advantage of component models is that they can allow synthesis when networks would otherwise be disconnected. However, in practise there is insufficient evidence to estimate interaction effects. Additive models should only be used where there is clinical plausibility (and preferably empirical evidence) that there are unlikely to be interactions.

2.4.4 Recommendations

- Hierarchical monotonic models (Owen 2015) and dose-response MBNMA models (Mawdsely 2016) may be used to connect and strengthen evidence networks if sufficient evidence on multiple doses is available.
- Class effect models may be used to connect and strengthen evidence networks if sufficient evidence is available to estimate them. Sensitivity analysis to treating different treatments as distinct interventions should be presented if possible.
- Additive component models may be used when there is clinical plausibility that there are unlikely to be interactions between components.

2.4.5 Research Recommendations

- Research to explore the relative performance of Hierarchical monotonic models (Owen 2015) and dose-response modelling (Mawdsely 2016) in NICE TAs examples would be valuable. An early draft of a manuscript looking at the value of MBNMA models to connect networks is already prepared. Extending this to compare the methods with hierarchical monotonic models and using examples from NICE TAs would take between 3-6 months.
- The Evidence Synthesis TSD series needs updating to include dose, class and component models. This would take approximately 2 months.

2.5 Bias adjustment

Key references: Dias et al. (2018), (Dias et al., 2011a) - TSD3

2.5.1 When the methods are likely to be useful

Bias adjustment methods may be useful when there are concerns over the methodological quality of RCTs included in a synthesis. The concern is that there are trial-level variables (e.g. related to study conduct) which modify the relative treatment effect, and have not been accounted for, thus potentially affecting the relative effect estimates from the (network) meta-analysis, and/or manifest as excess heterogeneity and, in a NMA with loops of evidence, inconsistency.

Meta-epidemiological studies have provided some evidence that bias tends to be greatest in trials with subjectively measured outcomes (Savovic et al., 2012b, Savovic et al., 2012c, Savović et al., 2018) (Burch et al., 2008), suggesting that it could be particularly important to consider bias adjustment for these outcome types, although a more recent study did not replicate this finding (Moustgaard et al., 2020).

The evidence base for associations with other study characteristics is smaller and uncertain. For systematic overviews of meta-epidemiological evidence across a range of domains see (Berkman et al., 2014) and Page et al. (2016).

It has been suggested that trials sponsored by industry tend to favour the product of the sponsor (Gartlehner et al., 2010, Gartlehner and Fleg, 2010, Flacco et al., 2015). Naci et al. (2014) used a meta-regression approach (section 2.5.4) to explore the effects of trials with and without industry sponsorship, in a network meta-analysis of statins for Low-density lipoprotein (LDL) cholesterol reduction. They found no evidence of industry sponsor effects in trials of statins for LDL cholesterol reduction when the dose of stating given in each arm of each study was taken into account. Earlier work (Dias et al. (2010a) and Barden et al. (2006)) also found no evidence of sponsorship bias in meta-analyses of antidepressants and acute pain and migraine trials, respectively. We hypothesise that as long as differences between studies, such as treatment dosing, are appropriately accounted for in a NMA, there is no need to adjust for "sponsorship bias".

Bias adjustment methods aim to transform estimates of treatment effect thought to be biased, into unbiased estimates which are then pooled. Bias adjustment is appropriate when some of the evidence provides potentially biased estimates of the target parameter. The meta-regression methods detailed below require a relatively large number of studies to be included in the synthesis, as well as a mix of studies considered at risk of bias and studies considered to provide unbiased estimates. Therefore, their applicability to the technology appraisal scenario, where often only a few studies are synthesised, may be limited. Using expert elicitation (section 2.4.2) or empirically-based prior distributions (section 2.5.3) are the methods most likely to be applicable to TAs. Accounting for missing outcome data (section

2.5.5) may also have a role where studies have a large proportion of missing outcome information.

2.5.2 Expert elicitation of bias distributions

Turner et al. (2009) proposed a method to replace a potentially biased study estimate with an adjusted estimate based on expert opinion.

Each study is considered by several independent experts who are asked to provide information on their understanding of each study's departures from an idealised protocol, which is then used to develop a bias distribution. A study can suffer from both internal *and* external bias and both can be adjusted for. The bias information on each study provided by each expert is combined into a single bias distribution. Multiple experts' distributions are pooled mathematically (O'Hagan et al., 2006, Turner et al., 2009) to create a new, adjusted, estimate of the treatment effect in that study (and its variance). The adjusted treatment effects for each study are then treated as the data inputs for a standard pairwise meta-analysis, indirect comparison or network synthesis.

This method can be used when the number of trials is small and meta-regression approaches (section 2.5.4) would fail. However, it is difficult and time-consuming to carry out in practice, experts find the required tasks challenging, and it is vulnerable to the subjectivity of the experts' opinions.

2.5.3 Meta-epidemiological sources for bias distributions

An alternative approach is to use results from collections of previous meta-analyses to provide empirically based prior information on the bias parameters in a new meta-analysis (Welton et al., 2009a). Savovic et al. (Savovic et al., 2012b, Savovic et al., 2012c, Savović et al., 2018) analysed data from meta-epidemiological databases, to obtain empirically based prior information for binary outcome data modelled on a log-odds scale and stratified by risk of bias indicator and outcome type. This prior information allows the analysis of a new meta-analysis to borrow strength from the studies at high risk of bias, whilst simultaneously adjusting for and down-weighting the evidence from those studies based on the empirical evidence.

However, this type of analysis assumes that the study-specific biases in the dataset currently being synthesised can be considered exchangeable with (i.e. similar to) those in the meta-epidemiological data used to provide the prior distributions used for adjustment (Welton et al., 2009a). As the degree of bias may be dependent on the type of outcome measure and vary by clinical area, different sets of prior distributions need to be selected and tailored for each problem (Savovic et al., 2012a). There has also been some work on how multiple indicators of risk of bias might interact (Savovic et al., 2012a), which suggest that the effects may be less than additive, although this result is very uncertain.

Currently there are no empirically based prior distributions for non-binary outcomes or for outcomes analysed on scales other than the log-odds ratio.

Prior distributions from meta-epidemiological studies can also be applied to NMA. However it is important to carefully define the direction in which bias is expected to act in studies comparing active treatments, otherwise biases between studies may “cancel out” and the mean bias will be under-estimated (Dias et al., 2010c). The meta-epidemiological evidence (Savovic et al., 2012b) excluded trials where it was not clear which direction bias would act. Chaimani et al. (2013) report results from a network meta-epidemiological study, where a collection of network meta-analyses were analysed to estimate bias resulting from indicators of risk of bias. However, this was restricted to networks where all treatments have been compared with a common comparator and where the direction of any potential bias was clear.

2.5.4 Estimation of bias by meta-regression

The methods presented in this section are based on “between-studies” comparisons and they provide no direct evidence for a “causal” link between the markers of study quality and the size of the effect. It is therefore important to establish that the results are statistically robust, and not dependant on a small number of studies.

2.5.4.1 In Network meta-analysis

The consistency assumption used in NMA means that, for networks with loops of evidence and sufficient numbers of studies at high and low risk of bias (on treatment comparisons in loops), there is redundant information which makes it possible to estimate bias parameters without requiring the strong “exchangeability” assumptions needed to adjust for bias using meta-epidemiological data (section 2.5.3) (Dias et al., 2010c). Assuming that the mean and variance of the study-specific biases are the same for each treatment comparison, it is possible to simultaneously estimate the treatment and bias effects in a single analysis (Dias et al., 2010c). Plausible assumptions for the direction of bias in active-active studies should be considered (Salanti et al., 2010, Dias et al., 2010c).

Dias et al. (2010c) present bias adjustment models in a network meta-analysis of fluoride therapies to prevent the development of caries in children. Salanti et al. (2010) applied a similar model to three network meta-analyses of chemotherapy and other non-hormonal systemic treatments for cancer (ovarian, colorectal, and breast cancer), assuming that there was a “novel agent” bias which was exchangeable across cancers. Naci et al. (2014) used this approach to assess sponsorship bias in a network meta-analysis of statins for Low-density lipoprotein (LDL) cholesterol reduction.

2.5.4.2 Application to small study/publication bias

Meta-regression models can also be applied to adjust for “small-study bias”, where it is believed that the smaller the study the greater the bias. Possible mechanisms for small study biases are publication bias, or the possibility that smaller studies were conducted under less rigorous conditions (eg without blinding or allocation concealment). The underlying assumption is that relative treatment effects are over-estimated in smaller studies and closer to the “true” effect in larger studies. The “true” treatment effect is then conceived as the effect that would be obtained in a study of infinite size, taken to be the intercept in a regression of the treatment effect against the study variance (Moreno et al., 2009b). Because this approach extrapolates the relative treatment effect to a study of infinite size, it requires that a certain number of relatively large studies are available. If available studies only have small sample sizes, results will be misleading.

In a pairwise meta-analysis of anti-depressants, Moreno et al. (2009a), Moreno et al. (2009b) show that the bias-adjusted estimate from this approach closely approximates the results found in a simple meta-analysis based on a register of prospectively reported data.

In NMA, appropriate assumptions would need to be considered for the direction of small-study bias in “active-active” studies (Salanti et al., 2010, Dias et al., 2010c). The NICE guideline for eating disorders (National Institute for Health and Care Excellence, 2017) used this method to adjust for small study effects in a NMA.

2.5.5 Other methods

2.5.5.1 Bias due to missing outcome data

Missing outcome data is common in RCTs and can occur for a variety of reasons, many of which can lead to biased effect estimates if not adjusted for. The challenge in pairwise and network meta-analysis is that we usually only have summary level data available, and so cannot use imputation methods at the individual level. If all of the RCTs included in the meta-analysis have reported effect estimates that are appropriately adjusted for missing data, then these can be combined in meta-analysis. However, when the attrition rates of primary studies depend on the size of the underlying treatment effect, results of a random effects meta-analysis will still be biased (Yuan and Little, 2009). In this case, it is necessary to account for the potential bias due to missing data in the synthesis.

Similarly to the bias-adjustment models above, this can be done by estimating or using informative prior distributions on parameters that define the missingness mechanism (i.e. how reasons for missing outcome data affect the relative treatment effect) and allowing this to adjust and down-weight the estimates from studies with a large proportion of missing outcomes.

This can be done in two stages (Higgins et al., 2008, White et al., 2008a) or in a single, joint analysis (Turner et al., 2015a, White et al., 2008b).

However, it is usually difficult to estimate missingness parameters within a meta-analysis because there is almost complete confounding between the random treatment effect and the random missingness elements. In fixed treatment effect models, particularly when some trials have only small amounts of missing data, then the data are sufficient to identify missingness parameters and "learning" can take place (White et al., 2008b, Spineli et al., 2013).

Methods have been proposed for binary data analysed on the log-odds scale and continuous (normal) data.

Turner et al. (2015a) present a general framework for a Bayesian estimation of the missingness parameter that allows for different parameterisations of the missingness parameter, facilitating the use of informative prior distributions. Turner et al. (2015a) applied this to pairwise meta-analysis and Spineli (2019), Spineli et al. (2019) extended the method to NMA. In an NMA there is a greater potential to learn about the missingness parameter, because of the "spare" degrees of freedom generated by the consistency equations.

In the absence of any prior information on the missingness mechanism, it is still important to reflect the additional uncertainty in effect estimates as a result of the missing data. Turner et al. (2015a) use flat prior distributions on the probability of an event in the missing individuals to reflect uncertainty due to missingness. This is preferable to down-weighting studies or imputing "best-case" and "worst-case" datasets which produces estimates of treatment effect that are artificially precise. Modelling the missingness mechanism allows propagation of the uncertainty to the estimated relative treatment effects naturally increasing uncertainty around the treatment effects, which can lead to lower between-study heterogeneity (indicating that some heterogeneity has been explained after adjusting for missing data) (Spineli, 2019). In addition, by modelling missing outcome data in an NMA, we are able to learn about the missingness mechanisms in each intervention (Spineli, 2019, Turner et al., 2015a).

Mavridis et al. (2014) proposed similar models for pairwise and network meta-analysis of continuous data, using a two-stage estimation procedure which estimates a missingness parameter. Mavridis et al. (2019) extended this to allow for adjustment in the presence of both missing and last observation carried forward (LOCF)-imputed outcome data, to estimate the treatment effect if complete follow-up was obtained.

2.5.6 Recommendations

- When there are concerns with the methodological quality (or size) of included studies consider bias-adjustment methods as follows:
  a) Methods for bias-adjustment using expert elicitation or empirically-based prior distributions for the bias parameters should be considered ***when the number of studies is small***.
  b) Meta-regression approaches, including adjustment for small study effects, can be considered in NMA ***where there are many studies and they are at varying risk of bias*** (for small study effects, small and large studies are required), and the direction of potential bias is clear. Adjustment for small study effects can also be used in pairwise meta-analysis (small and large studies are required).
  c) Methods to account for bias and additional uncertainty due to missing outcome data should be considered when the proportion of missing outcomes is large and related to the size of the relative treatment effects, and ***there are several studies at varying levels of missingness***.

2.6 Combining randomised and non-randomised evidence

2.6.1 When the methods are likely to be useful

When the evidence available from RCTs is sparse, it can be useful to consider using comparative non-randomised (observational) evidence to increase precision of relative effect estimates (Bartlett et al., 2019) or to connect otherwise disconnected networks of treatments. However, there are differences in magnitude of effect size between RCTs and non-randomised studies (Ioannidis et al., 2001) and the additional biases, likely to be present in observational evidence, need to be taken into account as otherwise there is a danger of weakening, rather than strengthening inferences.

A number of methods have been used to combine evidence from different sources, which include naïve pooling, inclusion of external sources of evidence as prior information and hierarchical modelling. These methods were originally introduced in standard pairwise meta-analysis and later generalised to NMA (Schmitz et al., 2013).

2.6.2 Observational studies to inform prior distributions

Key References: (Efthimiou et al., 2017, Schmitz et al., 2013)

Observational data can be used to construct prior distributions for the relative effect parameters in a Bayesian framework. This approach consists of two steps. In the first step, a meta-analysis/network meta-analysis is conducted on the non-randomised evidence to estimate the mean relative treatment effects for (some or all) basic parameters. The posterior distributions generated from the first step can then be used directly as prior distributions for the basic parameters of the NMA for randomised data in the second step, or these can be down-weighted by using an increased variance for the prior distribution (Sutton and Abrams, 2001). Alternatively, the predictive distribution can be used as long as there is sufficient observational evidence to reliably estimate the heterogeneity. However, this approach takes the observational data at 'face value' and does not adjust for any potential bias in the estimated mean effects from the observational data.

The posterior distributions estimated in the first step can also be adjusted to account for bias *(Efthimiou et al., 2017)* – see Section 2.5.5.

Limitations:

Random effects meta-analysis cannot be used to formulate predictive prior distributions when there are insufficient non-randomised studies, unless strong assumptions are made about heterogeneity *(Efthimiou et al., 2017)*. As randomised and non-randomised trials are analysed separately, the between-design variability is ignored (Schmitz et al., 2013).

These methods have been proposed as sensitivity analyses, but it is unclear how to choose the degree of down-weighting to be used in a bias-adjusted base-case analysis.

2.6.3 Power priors

Key References: Banbeta et al. (2019), Ibrahim and Chen (2000), Hong et al. (2018)

Adaptively informative priors have been proposed. A 'power transform prior' approach takes into account the differences in study design between the RCTs and the observational studies (Ibrahim and Chen, 2000). This approach down-weights the observational data so that they contribute less compared to data obtained from the RCTs. A down-weighting factor which varies between zero and one, with zero meaning that the observational evidence is entirely discounted, and one indicating that it is considered at 'face-value'. The power prior may be pre-defined and fixed, or estimated from the data. In the latter case, a modified power prior should be used (Duan et al., 2006).

The impact of different levels of weighting on the results of the NMA is assessed by considering a series of values for the down-weighting factor. This approach can mitigate bias in the observational data, but does not correct for it.

Commensurate priors were introduced to include historical trials into current clinical trials (Hobbs et al., 2011, Hobbs et al., 2013, Hobbs et al., 2012) but can be extended to combine non-randomised and randomised evidence. A hierarchical model is specified where the commensurability parameter in the prior controls the extent of the influence of non-randomised data in a meta-analysis. Hong et al. (2018) recommend that power priors and commensurate priors approaches are used together and the results should be compared.

Banbeta et al. (2019) develop modified power priors which allow for the possibility of conflict between the observational and RCT data in the context of combining historical trials with current clinical trials.

Limitations:

It is unclear how to choose the degree of down-weighting to be used in a bias-adjusted base-case analysis and therefore multiple sensitivity analyses would need to be conducted.

2.6.4 Mixture priors

Key References: Röver et al. (2019)

In this approach heavy-tailed mixture priors are used to make use of external evidence when data are sparse. This is effectively a model-averaging approach where conditional posterior distributions are computed separately under two models, one for the randomised data and a second for the non-randomised data. These partial results are then recombined using Bayes factors (Röver et al., 2019). These prior distributions are robust to prior/data conflicts. Mixture priors simplify computations as off-the-shelf software can be used to perform the main computations, which then only need to be recombined.

2.6.5 Hierarchical modelling

Key References: (Efthimiou et al., 2017, Schmitz et al., 2013, Prevost et al., 2000)

Another approach allowing to differentiate between study designs is to introduce another level in the Bayesian hierarchical model for the NMA, to model the between-studies heterogeneity of treatment effects within each study design (RCT or observational) and across study designs. The design-specific summary estimates are then pooled in a joint network meta-analysis by assuming that they are exchangeable *(Efthimiou et al., 2017, Schmitz et al., 2013, Prevost et al., 2000)*.

Hierarchical models allow for adjustments to be made to account for systematic bias and for trials to be weighted according to design. Overall estimates can be compared to estimates at study design level to ensure consistency. Hierarchical models also allow for between-design heterogeneity to be accounted for (Schmitz et al., 2013).

This model can be extended by combining the power prior method described in Section 2.6.3 with the hierarchal model in order to provide a further sensitivity analysis. This can be achieved by introducing a multiplicative factor to the variance for the observational data (Schmitz et al., 2013).

The level of uncertainty (in terms of the credible intervals) is generally greater when using the hierarchical model compared to other approaches, since it explicitly accounts for the differences in study designs, thus allowing for additional variability across studies.

2.6.6 Other methods

2.6.6.1 Naïve pooling

Naïve pooling simply takes the observational data at 'face-value' and pools them with the RCT data. That is, data are included in the meta-analysis regardless of the study design. This does not account for the differences between the designs of the studies particularly in terms of the potential additional bias from observational evidence and its additional precision, as typically non-randomised studies are larger and therefore give more precise estimates of effect than RCTs. Therefore, there is a high risk of allowing the biases in the observational data to dominate the pooled result.

Naïve pooling can also help close loops in networks, allowing the estimation of consistency between direct (e.g. observational) and indirect (e.g. RCT) evidence (Schmitz et al., 2013). However, it is unclear what the implications of finding, or not finding, evidence of inconsistency would be. Should inconsistency be found, the RCT evidence should be preferred and adding the observational evidence is of no value; if inconsistency is not found, we cannot conclude that there is no inconsistency since currently available methods for inconsistency checking lack power to detect it. Thus, naïvely including observational evidence in an RCT meta-analysis has only limited, exploratory value to provide an initial insight into the effect of including non-randomised studies *(Efthimiou et al., 2017)*.

Limitations:

Naïve pooling assumes that there are no differences between different trial designs. It does not allow for any bias adjustments or to account for any additional uncertainties.

2.6.6.2 Design adjusted analysis

Design adjusted analysis is an extension of the naïve pooling approach, where studies are treated differently depending on trial design. In this method, the mean effect sizes and/or variances of non-randomised studies are adjusted before synthesising them with randomised studies. Mean effect sizes are shifted by a bias term and variance is inflated such that the weight of the study is decreased. This is similar to the bias-adjustment methods described in section 1.5.

This method assumes that the observational data estimate a biased effect, but that the bias parameter can be independently estimated or prior distributions on the extent and direction of bias can be elicited from experts(Turner et al., 2009, Schnell-Inderst et al., 2017), so that the effect estimate arising from the observational data can be adjusted and down (Efthimiou et al., 2017, Verde and Ohmann, 2015). However, defining the magnitude and direction of the bias terms a priori can be challenging as the bias in estimates of relative effects from non-randomised studies could depend on many factors. Currently there is limited meta-epidemiological evidence to inform this (Anglemyer et al., 2014).

2.6.7 Recommendations

- Careful consideration of whether the observational data are sufficiently credible and how the results should be interpreted is required. A bias-adjusted base-case should be used with other methods considered as sensitivity analyses.
- Methods that attempt to down-weight and adjust the observational evidence prior to inclusion in the synthesis are preferred (hierarchical model, design-adjusted analysis).
- Naïve pooling of randomised and non-randomised evidence is not recommended, although it may be useful as a first step analysis, or as a sensitivity analysis.

2.6.8 Research Recommendations

- An MRC funded project is underway to describe the circumstances under which simultaneous bias estimation and adjustment can be carried out, and when conflict between prior distributions on the bias and the data can be detected. This will use case-studies to demonstrate potential benefits of bias-adjustment approaches and to estimate the degree of bias in observational evidence compared to RCT evidence.

As well as academic publications, a TSD could be produced outlining the specific implications for TA.

3 MULTIPLE OUTCOMES

When conducting evidence synthesis of existing studies to inform HTA decision-making, relevant studies may not provide direct evidence about all the outcomes of interest. This may be due to, for example, differences in measurement or reporting between studies (e.g. different scales used or different follow-up times), outcome reporting bias or early licensing of new health technologies evaluated based on a surrogate marker when data on the final clinical outcome are not yet available, but rapid approval processes are desirable in particular in areas of highest priority in health care (such as cancer). Studies that do not provide direct evidence about a particular outcome or treatment of interest are often excluded from a meta-analysis evaluating that outcome or treatment and, hence, data from patients in those studies would not contribute to HTA decision-making process. This is undesirable, especially if the number of studies for new health technologies is limited and also if the study participants are otherwise representative of the population, clinical settings and condition of interest. In case of data limited to surrogate markers, not including them would lead to substantial delays in HTA policy decisions. Statistical models for multivariate meta-analysis address these challenges by simultaneously analysing multiple outcomes.

This section discusses multivariate normal random effects (MVNRE) meta-analysis, and a models for structurally related outcomes. The MVNRE approach assumes nothing about the outcomes, except that they are correlated at both the within-trial (individual) level, and at the between-trial level. Structural models seek to capture clinical, or logical, relationships at either the individual or trial level, or at both levels.

The purpose of analysing separate outcomes simultaneously, within a single model, is threefold:

1. It should generate more precise estimates of treatment effects, or at least more robust, as they will be based on more data.
2. In most applications NMAs on single outcomes will not be fully connected, while models synthesising over multiple outcomes are more likely to be connected, and to be much richer in connections, contributing further to precision, robustness, as well as opportunities to check consistency and underlying structural assumptions.
3. If more than one of the outcomes appears in the CEA, it is essential that correlations between outcomes are correctly propagated; otherwise CEACs will be incorrectly calculated, often seriously so.

These issues are detailed in the sections below.

3.1 Multi-variate Normal Random Effects

Key references: (Bujkiewicz et al., 2019a, Riley et al., 2007, Jackson et al., 2011)

3.1.1 When the methods are likely to be useful

3.1.1.1 Borrowing of information to increase precision

Borrowing of information across outcomes may be of interest to analysts and will be mostly effective, in terms of the increased precision, when the outcomes are highly correlated and the percentage of studies with missing outcomes is large (Riley et al., 2017). It is difficult to state the minimum/maximum number of studies as this will depend on the magnitude of the correlation: the level of borrowing of strength will be less when correlation is low compared to meta-analysis of the same number of studies for each outcome and stronger between-studies correlation.

In the case of *multivariate network meta-analysis*, borrowing of strength will result from both sharing information across treatments through indirect effects and across outcomes. Often multivariate NMA will not improve the precision further. However, as shown in an example of joint modelling of treatment effects on relapse rate and adverse events in multiple sclerosis in TSD20, borrowing of information across outcomes can be observed in NMA when data on one (or subset) of the outcomes are particularly sparse.

3.1.1.2 Increased evidence base

Regardless of whether or not the analysis leads to improved precision of the estimates compared to the univariate case, it may still be valuable to carry out due to an increased evidence base to all relevant studies. Discarding studies not reporting primary outcome of interest may be considered research waste, especially if the study participants are otherwise representative of the population, clinical settings and condition of interest. This can also increase the representativeness of evidence base.

3.1.1.3 Potential for reduced impact of outcome reporting bias

Multivariate meta-analysis may still be valuable (even if no increase in precision is detected) as it provides a sensitivity analysis for potential impact of outcome reporting bias(Kirkham et al., 2012). That is, by borrowing information from correlated outcomes that are reported, the analyst can reduce the impact of unreported outcomes, and see if conclusions are robust. More recently, it has been shown that the impact of outcome reporting bias is also reduced by use of multivariate NMA(Hwang and DeSantis, 2018).

3.1.1.4 Appropriately considering natural correlations in decision modelling

A health-economic model typically uses inputs from multiple outcomes which are propagated through the model. Using a multivariate posterior distribution and accounting for the correlation between the outcomes may reduce bias of the resulting cost-effectiveness estimates.

When multiple clinical endpoints are included in an economic model, the multivariate meta-analysis accounting for the correlation between these endpoints will be important for two reasons: (1) it may change the point estimate of the incremental cost-effectiveness ratio (ICER) if the model includes a non-linear function of the two endpoints (these ICERs will in fact be biased unless the correlation is reflected); and (2) it may change the estimates of decision uncertainty generated by the models i.e. probabilities that interventions are cost-effective or value of information estimates. Taking account of the correlation between multiple endpoints could, therefore, have important implications for decision making, as many models utilise two or more decision endpoints, such as PFS and OS in cancer models (Woods et al., 2017).

3.1.2 Bivariate random effects meta-analysis

Key references: (Bujkiewicz et al., 2019a, Bujkiewicz et al., 2019b, Riley et al., 2007, Jackson et al., 2011) (Bujkiewicz et al., 2017)

A general form for the bivariate random effect meta-analysis with bivariate normal random effects has been described by a number of authors *(Bujkiewicz et al., 2019a, Bujkiewicz et al., 2019b, Riley et al., 2007, Jackson et al., 2011)*. In a Bayesian framework, prior distributions are placed on the mean effects and the elements of the between-studies covariance metrics (the heterogeneity parameters for the two outcomes and the between-studies correlation). The method can be implemented using the standard form with bivariate normal random effects (Bujkiewicz et al., 2019a) or in the product normal formulation (Bujkiewicz et al., 2017). Both approaches are covered in detail in TSD20. The approaches are equivalent when there is no missing data, but can give slightly different results in the presence of missing outcome data. When the strength of the association between the treatment effects on the two outcomes is of interest (as it typically is in the context of surrogate endpoints, as discussed in section 3), product normal formulation gives more detailed information (in the form of the intercept, slope and conditional variance) compared to the standard form which only provides the correlation. However, all the parameters (such as correlations or slopes) can be derived from the parameters produced by each method.

3.1.2.1 Examples of multivariate meta-analysis used in the context of NICE

Bivariate meta-analysis has been used in a technology appraisal of *continuous positive airway pressure devices for the treatment of obstructive sleep apnoea-hypopnoea syndrome* conducted by NICE (TA139 https://www.nice.org.uk/guidance/ta139), where modelling two endpoints jointly in a bivariate meta-analysis allowed the analysts to include all of the studies in the meta-analysis. In this case results were not dramatically different from those obtained using conventional univariate meta-analysis, however such analysis is still valuable as it provides a sensitivity analysis for potential impact of outcome reporting bias.

Another NICE technology appraisal (TA383) that used bivariate meta-analysis was for *tumour necrosis factor-α inhibitors for ankylosing spondylitis and non-radiographic axial spondyloarthritis* (https://www.nice.org.uk/guidance/ta383). A decision model was developed with a generalised framework for evidence synthesis of two outcomes to determine the long-term quality-adjusted life-years and cost burden of the disease in the economic model. The bivariate approach not only allowed all relevant evidence to contribute to the synthesis but also ensured that all measures were synthesised together to reflect the expected correlations between the effects on the two outcomes. Uncertainty was also more appropriately quantified compared to synthesising each outcome separately.

Both examples are summarised in more detail in TSD 20.

3.1.3 Multivariate random effects meta-analysis

Key references: (Wei and Higgins 2013b, Achana et al., 2014, Bujkiewicz et al., 2016, Bujkiewicz et al., 2013)

Compared to the bivariate meta-analysis, there is an additional complexity in multivariate meta-analysis of more than two outcomes, in particular when modelling them in a Bayesian framework.

3.1.3.1 Models with prior distribution on the covariance matrix

A prior distribution can be placed on the whole covariance matrix to ensure the matrix satisfies the conditions to be a proper covariance matrix. There are a number of approaches to constructing a suitable prior distribution on the between-studies covariance matrix, for example by using so called separation strategy with either Cholesky or spherical decomposition (Wei and Higgins 2013b), which both are appropriate. When Cholesky decomposition is used to construct a prior distribution, results may depend on the ordering of the outcomes (in the way they are stored in the data and enter the analysis) and a sensitivity analysis is recommended.

Another option which has been explored (Wei and Higgins 2013b) is the use of the inverse Wishart prior, however this approach has been shown to be unreliable (Wei and Higgins 2013b), (Bujkiewicz et al., 2017), as it can be very influential on the posterior estimates of the between-study variances.

3.1.3.2 Models using product normal formulation

An alternative modelling approach has been proposed to describe the between-studies variability in the product normal formulation, which replaces a multivariate matrix form of the model with the series of univariate conditional distributions (Bujkiewicz et al., 2013). It is a convenient approach which simplifies the model by assuming that treatment effects are conditionally independent between some of the outcomes, which can help reduce the number of parameters to estimate. However, the full correlation structure, assuming that the treatment effects on all the outcomes are correlated, is also possible as well as assuming different forms of structures for the correlation matrix (deciding which pairs of outcomes are correlated or conditionally independent) (Bujkiewicz et al., 2016). Derived relationships between the parameters of the model ensure proper definition of the covariance structure.

3.1.4 Informative vs non-informative prior distributions for the correlations

3.1.4.1 Informative prior distribution

Informative prior distribution on the between-studies correlation can improve borrowing of information. Such prior distributions can be constructed by performing multivariate meta-analysis of external data – perhaps of the same treatment(s) but slightly different population as in an example in TSD20 (using a prior for the correlation based on data on first line treatment in the multivariate meta-analysis of second line of therapies (Bujkiewicz et al., 2013) – only prior distributions for the correlations would be appropriate to use, not prior distributions for the treatment effects).

- *Weakly informative prior distributions*

Note that weakly informative prior distribution(s) on the between-studies correlation(s), for example assuming positive or negative correlation (distributions restricted to values between 0 and +1 or between -1 and 0 respectively), are more appropriate – the seemingly non-informative flat prior distributions on the correlation, such as uniform prior ranging between values of -1 and +1, can results in biased pooled estimates (Burke et al., 2018).

3.1.5 Methods for dealing with unreported within-study correlations

Key references: (Riley, 2009, Riley et al., 2014, Wei and Higgins 2013a, Bujkiewicz et al., 2013)

When using multivariate meta-analysis, analysts may come across issues related to the estimates of the within-study correlations being unavailable. Ishak et al (Ishak et al., 2008) and Riley (Riley, 2009) highlighted the importance of accounting for the within-study correlation and the dangers of ignoring it (by for example setting it to zero). The within-study correlation (between treatment effects on multiple outcomes) is assumed known in the multivariate meta-analytic methods, however typically it will not be reported by the original articles (the correlation can be, for example, between log odds ratios for multiple dichotomous outcomes and, as such, is unlikely to be reported).

3.1.5.1 Using IPD directly or via bootstrapping

Individual participant data (IPD) are needed to estimate the within-study correlation(s). Such estimation can be done by either modelling the treatment effects on the multiple outcomes jointly (Riley et al., 2015) or by bootstrapping (Daniels and Hughes, 1997) to generate multiple treatment effects on two (or more) outcomes and then taking the correlation between them. Alternatively, double bootstrapping can be carried out to obtain the correlation with uncertainty. When this is carried out on IPD from a source external to the meta-analysis, the double bootstrap technique can be used to construct an informative prior distribution for the within-study correlation(s) (Bujkiewicz et al., 2014, Bujkiewicz et al., 2013).

3.1.5.2 Approximation methods

Alternative methods have been developed providing approximations to the between-studies correlations between treatment effects on more complex scales (e.g. log ORs) to correlations between, for example, probabilities of event (Wei and Higgins 2013a), which may be reported by some studies.

3.1.5.3 Alternative model by Riley et al

An alternative formulation of bivariate meta-analysis for studies with unknown within-study correlation can be implemented, which combines covariances from both the within-study and the between-study model in a single term (Riley et al., 2008). The limitation of the method is that it won’t provide an estimate of the between-studies correlation, if such estimate is of interest.

3.1.5.4 Sensitivity analyses

When no IPD or approximate measures are available, sensitivity analyses can be carried out using a range of values for the within-study correlations to investigate the impact of their magnitude on the results. This can help assess the feasibility of an analysis using the multivariate approach.

3.1.6 Methods for multivariate network meta-analysis

Key references: (Achana et al., 2014, Efthimiou et al., 2014, Hong et al., 2015, Bujkiewicz et al., 2019a, Jackson et al., 2018)

3.1.6.1 Contrast based methods with arm-level data entry

Achana et al. developed a model for network meta-analysis of treatment effects on multiple correlated outcomes in multi-arm studies (Achana et al., 2014). The authors extend the standard NMA model (contrast based with arm-level data entry, such as mean effect or log odds of an event in each arm) to multiple outcome settings in two stages. In the first stage/model, information is borrowed across outcomes through modelling the within-study and between-studies correlation structure, accounting for multi-arm studies and allowing for any number of outcomes. It is assumed that at the within-study level the estimates of treatment effects in each arm on multiple outcomes follow multivariate normal distribution (of a dimension equal to the number of outcomes) for each study. At the between-studies level, the true treatment effects follow a multivariate normal distribution common to all studies of the same treatment contrast. The method is then extended to multi-arm trials. In the second stage/model, an additional assumption is made, of equal or constant relative potency of treatments across outcomes which imply exchangeability of the relative effects between the non-reference/baseline treatments (by expressing basic parameters as a sum of treatment-specific effects and outcome-specific effects). This enables prediction of treatment effects on outcomes for which evidence is either sparse or the treatment effects had not been considered by any one of the studies included in the analysis. Applicability of this final model may be limited, to, for example, composite interventions, as in the public health data example used by the authors where different types of service provision in addition to educational interventions have been investigated. Software implementation assumes homogeneity of the between-studies correlations and heterogeneity parameters across treatment contrasts.

3.1.6.2 Contrast based methods with contrast-level data entry

Efthimiou et al. proposed a model for the joint modelling of multiple endpoints at contrast level data entry, where data are treatment differences, and specifically log odds ratios (Efthimiou et al., 2014). In contrast to the arm-level data entry model by Achana et al, at the within-study level the estimates of treatment differences (rather than treatment effects in the arms) follow a multivariate normal distribution across all outcomes. The authors have focused largely on the bivariate case, but also allow for multi-arm trials (Efthimiou et al., 2014). Efthimiou et al. then developed an alternative model which is a network extension of an alternative multivariate meta-analytic model (Riley et al., 2008) combining the within-study and between-studies covariance terms into a single term, avoiding the need to specify the within-study correlations (Efthimiou et al., 2015).

Bujkiewicz et al implement bivariate NMA models (Bujkiewicz et al., 2019a). Their models are limited to two-arm studies and two outcomes, but allow for heterogeneous between-studies correlations and heterogeneity parameters across the treatment contrasts. The models are largely developed to investigate surrogacy patterns between treatment effects on two outcomes, when such patterns can vary between the treatment contrasts. However, they can equally be used for the purpose of obtaining means and indirect effects (results reported in supplementary materials to the paper). In some applications, using heterogeneous between-study covariances may lead to more appropriate results. The models extend with the additional assumption that pooled treatment effects on the two outcomes in each study arm are exchangeable across treatment arms, to allow for predicting effects for all outcomes, when some outcome-treatment combinations are unreported and not available through indirect comparisons. This can be particularly advantageous when data are sparse or when a new study investigating a new treatment reports the effect of the treatment only on a subset of outcomes. However, the exchangeability assumption is strong and has to be considered carefully.

3.1.6.3 Contrast based methods with contrast-level data entry and inconsistency term

Jackson et al derived a frequentist method of moments for synthesis of data on multiple treatments and outcomes in a multivariate NMA allowing for inconsistency, using design-by-treatment interaction (Jackson et al., 2018). Their implementation of the method in R allows for both estimations: using a model with assumption of consistency in the network and a model with the inconsistency term, which can be a useful tool for assessing inconsistency in multivariate network meta-analysis. However, we do not recommend inconsistency models for use in decision-making.

3.1.6.4 Arm-based methods

Hong et al developed multivariate methods for IPD NMA (Hong et al., 2015) and for multivariate NMA with missing data (Hong et al., 2016a) using two approaches: contrast based approach and arm-based approach. Arm-based approaches to NMA have been a topic of a debate in terms of the appropriateness of the methods. This is discussed in Section 1.1.2 of this report.

3.1.7 Computational issues

Multivariate meta-analysis, in particular in larger dimensions beyond the bivariate case, can be computationally difficult. TSD 20 provides tools for implementing pairwise multivariate meta-analysis of any number of outcomes using separation strategies, methods using product normal formulation are available for trivariate case and some information is available how to extend the methods to more outcomes than three. These methods are easier when assuming some structure on the covariance matrix. TSD 20 also covers methods for multivariate NMA for any number of outcomes and multi-arm trials.

3.1.8 Recommendations

Multivariate meta-analysis is could be considered in a number of situations.

- When number of studies in meta-analysis reporting an outcome of interest is small, and combining the analysis with other outcomes in a multivariate framework increases the evidence base if some additional studies report other outcome(s), but not the outcome of primary interest. This is likely to be effective if the between-studies correlation is high (and the precision can be further improved by the use of an informative prior distribution on the correlation).
- Sensitivity analyses with the outcomes analysed separately are recommended to allow the decision maker to qualitatively ascertain the extent of borrowing. It would be important that the evidence sets are described in detail so that potential imbalances are reflected in the interpretation of results.
    - Increased evidence base can still be of value even if the precision of the average effect is not improved. Discarding the studies that do not report the outcome of interest but otherwise are representative of the scope of the HTA can be difficult to justify. Research studies require considerable costs

and time, and involve valuable patient participation, and simply discarding them could be viewed as research waste.

- Use of multivariate meta-analysis can produce more robust estimates in the presence of suspected outcome reporting bias.
- If a decision model includes parameters defined by two (or more) outcomes, the joint posterior distribution can be propagated through the model allowing for more appropriate representation of uncertainty, taking into account the correlation, thus resulting in more appropriate estimates of cost-effectiveness and decision uncertainty.
- When risk-benefit analysis is of interest, multivariate meta-analysis of effectiveness estimates and those of adverse events can result in reduced uncertainty around the safety outcomes (compared to a univariate meta-analysis of safety outcomes only), as trials are typically not powered to detect an impact of treatments on adverse events.

3.1.9 Research recommendations

- Further research on when the added complexity is worthwhile would be important to guide NICE's methods recommendations.
- An extension of TSD20 is required to cover multivariate NMA in more detail, and in particular provide code for implementation of wider range of methods – e.g. contrast based method for contrast-level entry data with multi-arm trials (the work is under way and the TSD extension will be feasible in 12-18 months).

3.2 Structural models

Key references: Dias et al. (2018), Pedder et al. (2019)

Structural models aim to take account of particular relationships between each outcome, are each tailored to specific clinical scenarios, and are therefore disease dependent. It is essential that their biological plausibility is checked by clinical experts, and where possible that the structural assumptions are checked statistically. Here, we focus on a set of examples most likely to be relevant in submissions to NICE appraisals.

We note that there are several “multiple outcome” models already in common use at NICE relating to ordered categories and competing risks which are covered in TSD2. These include ordered probit or ordered logit models used for PASI or ACR scores in psoriasis and rheumatoid arthritis respectively, models for mutually exclusive end-points, for example in trials of treatments for schizophrenia (Dias et al., 2013, Dias et al., 2011b), and models for treatment discontinuation and treatment efficacy conditional on continuation (Dias et al., 2018, National Clinical Guideline Centre, 2012, National Collaborating Centre for Mental Health, 2014).

It is worth noting that these structural models are motivated mainly by the lack of individual patient data from all the included trials. Simpler analyses would be possible if IPD were comprehensively available (section 2.1).

3.2.1 Multiple outcomes reported in different ways

Trials may report results in different ways, and some trials may report data in more than one way. For example in studies of treatments for influenza studies reported one or more of: mean (or median) time to end of fever; mean (or median) time to end of symptoms; proportion of patients reaching end of symptom or end of fever outcomes within X days. These can all be synthesised simultaneously through a model that expresses the various reported statistics as a function of underlying shared parameters (Welton et al., 2008, Burch et al., 2010). These clinical states have a fixed, known, temporal order within a natural history model. Modelling can take advantage of that known relationship and the jointly estimated treatment effects.

Other examples of outcomes reported in different ways and synthesised using a shared parameter model have been developed ((Keeney et al., 2018, Dias et al., 2011b)) and these could also be adapted to the multivariate case.

3.2.2 Simultaneous mapping and synthesis

The term “mapping” is often used in HTA to reflect the change of scale of the relative treatment effects obtained by multiplying a treatment difference on a disease-specific measurement scale by a “mapping coefficient” (Brazier et al., 2010).

Models for simultaneous mapping and synthesis have been developed, in which the mapping coefficients linking several test instruments are estimated from the RCT evidence making up the meta-analysis of interest. The model simultaneously synthesizes both the mapping coefficients and the treatment effect information within and between trials to produce estimates that are invertible and transitive (Lu et al., 2014).

These models are, in effect, special cases of MVNRE models (see section 2.1) in which it is assumed that the relative treatment effect on a number of outcomes is the same to within a "mapping coefficient". There is an implication that if there is no treatment effect on one outcome, there is no treatment effect on the others. The synthesis estimates both the relative treatment effect (on any of the outcomes), and a coefficient that maps from the relative treatment on a reference outcome onto treatment effects on all the other outcomes. This type of model has been used for outcomes in trials of treatments for depression, social anxiety, and ankylosing spondylitis (Lu et al., 2014, Ades et al., 2015, Kounali et al., 2016). It could probably be applied to a wider set of conditions, wherever the assumptions are clinically reasonable, although it requires a connected network of outcomes (i.e. that the combination of outcomes reported in the included RCTs form a connected network) (Dias et al., 2018).

3.2.3 Repeat Observations for continuous outcomes

Where trials report continuous outcomes at several points in time, multi-variate models should be fitted that take account of the correlation between time points.

If the relative effect is assumed to not change over time, only the correlation structure needs to be attended to. A univariate model can be used where the measures over time are pooled with the variance adjusted to take account of the correlation structure (National Institute for Health and Care Excellence, 2013).

If the treatment effect is assumed to change over time (clinical input should be sought) a suitable functional form can be estimated to represent the relationship between treatment effects over time. Jansen et al. (2015) proposed fitting fractional polynomials models. More recently a model based network meta-analysis (MBNMA) modelling framework has been proposed that allows for non-linear modelling of multi-parameter time-course functions, accounts for residual correlation between observations, preserves randomisation by modelling relative effects, and allows for testing of inconsistency between direct and indirect evidence on the time-course parameters (Pedder et al., 2019). The MBNMA framework can include a multivariate Normal likelihood, but rather than assuming a multivariate Normal distribution for relative effects (section 2.1.6), a more structured model it fitted to capture the time-course relationship, giving more precise and interpretable results. This approach can take advantage of additional knowledge on the time-course function for the onset of action of different treatments, from early pharmacokinetic studies, or when enough studies reporting at different time points are available, the different functional form can be assessed. An R package is available to fit time-course MBNMA models (https://cran.r-project.org/web/packages/MBNMAtime/index.html).

3.2.4 Repeat observations on binomial outcomes

Trials often report the proportion of patients reaching an end-point at more than one follow-up time, and different trials may report at different follow-up times. To synthesis all the data available, a piece-wise constant hazard model can be constructed, when the end-point is an absorbing state (eg mortality). The relative treatment effect can be allowed to vary between time intervals, or not. “Smoothing” of the baseline hazard rate and/or the hazard ratio across adjacent time intervals can be accomplished by random walk processes. These models have been applied to gastro-intestinal reflux disease (Lu et al., 2007) and trials of stents (Stettler et al., 2007, Stettler et al., 2008).

When the binary outcome may change over time for each patient (eg having a headache free day), then the MBNMA time-course modelling framework can also be applied to model relative effects as log-odds ratios as a function over time (Pedder et al., 2019).

3.2.5 Synthesis of Markov transition parameters

Occasionally trials report the number of transitions between health states. In the case of asthma, for example, these would be transitions between “successfully treated weeks”, “unsuccessfully treated weeks”, “exacerbation”, and “treatment failure” (an absorbing state). This can be represented as a Markov transition model, with a “baseline” and relative treatment effect for every transition. Alternatively, if clinically appropriate, the model can be cast as a Markov rate process with transitions only between adjacent states. This reduces the number of transitions, and makes it possible to identify the transition(s) on which the treatment operates. The model can be written as first order differential equations (Price et al., 2011, Price and Briggs, 2002). Data from trials reporting results at different time points could be incorporated within the same model (Welton and Ades, 2005). In NICE TAs, transitions between Markov model states are rarely synthesised, and instead individual patient data from the manufacturers trial used directly to estimate transition probabilities (for example the ongoing TA for Dupilumab for severe asthma with type 2 inflammation). Methods that synthesise evidence on transition rates to obtain rate ratios are desirable, as RCTs are designed to provide the best source of relative effects, and may not be the best source of absolute effects. Rate ratios from a synthesis of RCTs could be applied to transition rates on the reference treatment from a suitable representative prospective cohort study.

3.2.6 Recommendations

- The Methods Guide should be updated to the effect that methods for ordered category data, and for trials involving both discontinuation and efficacy outcomes, are well established, and should be used wherever appropriate.
- Joint synthesis of structurally related outcomes is recommended wherever possible, to increase precision and robustness of decision making. Models should be informed by knowledge of the natural history of the disease and checked for clinical plausibility. The underlying assumptions should be checked statistically wherever possible. Those making submissions can be guided by the above examples, but should consider whether structural models would contribute precision and robustness in other clinical areas too.

3.2.7 Research Recommendations

- There is a research need to explore the application of methods for synthesis of Markov transition rate parameters in NICE TA examples.
- There is a need for a TSD on evidence synthesis with structural models for multiple outcomes, including time-course modelling.

4 SURROGATE ENDPOINTS

Key references: (Bujkiewicz et al., 2019a, Daniels and Hughes, 1997)

Surrogate endpoints (such as, for example, progression free survival (PFS) as an early marker of overall survival (OS) in cancer) play an increasingly important role in the drug development process as new health technologies are increasingly being licensed by the regulatory agencies, such as European Medicines Agency (EMA) in Europe or Food and Drug Administration (FDA) in the US, based on evidence obtained by measuring effectiveness on a surrogate marker. When data on the final clinical outcome are not available or limited at the licensing stage, this will also be the case for the HTA decision-making process. To expedite availability of new therapies to patients, technology appraisals may need to be based on an estimate of the effect of the therapy measured on the surrogate endpoint.

HTA agencies, however, are cautious about the use of surrogate outcome data and highlight the importance of an appropriate use of such endpoints in their guidelines. This is particularly reflected in the guidelines published by the European Network for Health Technology Assessment (EUnetHTA) as well as NICE's current guidance on methods for manufacturers. The guidelines recommend that HTA analysts and decision-makers should be cautious about using surrogate endpoints and use them only if they have been appropriately validated. The additional uncertainty associated with using surrogate endpoints to predict cost-effectiveness should also be fully explored.

4.1 Validity of surrogacy

Before they can be used in evaluation of new health technologies, candidate surrogate endpoints have to be assessed for their predictive value of the treatment effect on the final clinical outcome. Relying solely on patient level association is not sufficient when evaluating surrogate endpoints, in particular when individual level association has been evaluated based on data from a single trial (Fleming and DeMets, 1996). A single trial validation cannot guarantee that an association between effects confirmed based on individual data under one treatment will hold in other interventions. A meta-analytic approach, based on data from a number of trials to establish the association between the treatment effects on the candidate surrogate endpoint and on the final outcome is more appropriate for evaluation of surrogate endpoints.

A modelling framework is required to establish the strength of the surrogate relationship between the treatment effects on the surrogate and the final outcome and to predict the likely treatment effect on the final outcome for the new health technology. Multivariate meta-analytic methods provide such a framework as they, by definition, take into account the correlation between the treatment effects on the surrogate and final outcomes as well as the uncertainty related to all parameters describing the surrogate relationship.

Validity based on a pre-specified criteria, however, may be arbitrary, and validation based on evaluating predictive value of the surrogate, which takes into account uncertainty associated with the strength of the surrogate relationship as well as around of the treatment effects in the data, is preferable.

In practice, it is difficult to quantify how large the correlation should be in order to consider the surrogate endpoint suitable to make the prediction. Some authors claimed that a high level of association is required to demonstrate surrogacy. For example, Lassere et al in their Biomarker-Surrogacy Evaluation Schema defined such high association by the square of the between-studies correlation (or so-called adjusted R-squared) above 0.6 (Lassere et al., 2012), and the German Institute of Quality and Efficiency in Health Care (IQWiG) requires high correlation with the lower limit of the 95% confidence interval above 0.85 (https://www.ncbi.nlm.nih.gov/books/nbk198799/). Other authors emphasised that the decision of whether the surrogate endpoint can be used to predict clinical benefit, should be based on the balance between the strength of the surrogate relationship and the need for the decision to be made about the effectiveness of the new treatment, for example for regulatory purposes (Alonso A, 2016).

As discussed by Bujkiewicz et al., the strength (or weakness) of the surrogate relationship will manifest itself in the width of the predicted interval of the treatment effect on the final outcome (Bujkiewicz et al., 2019b). A smaller value of the correlation will result in a larger interval and hence increased uncertainty about the regulatory or clinical decision made based on such prediction. The implication of this is that perhaps we don't need criteria about the correlation and instead we need only look at the predictions (Bujkiewicz et al., 2019b). Such predicted estimate (along with the uncertainty) of the treatment effect may be used in HTA decision making. The evaluation of the quality of predictions can be achieved through a cross-validation procedure (Daniels and Hughes, 1997) and TSD20 section 2.3.3 (Bujkiewicz et al., 2019a).

4.1.1 Aggregate data based methods for surrogate endpoint evaluation and predictions

Key references: (Daniels and Hughes, 1997, Bujkiewicz et al., 2017, Bujkiewicz et al., 2019a)

Daniels and Hughes proposed a model, referred to in TSD 20 as the "standard surrogacy model", which assumes that the estimates of the treatment effects on the surrogate and final outcomes follow bivariate normal distributions in each study (Daniels and Hughes, 1997). These effects can be, for example log hazard ratios for PFS and OS. At the between studies level, they assume fixed effects regression-like relationship, and the regression parameters describe the strength of the surrogate relationship. The same model can be used to make predictions of the (unreported) treatment effect on the final outcome from the treatment effect measured on the surrogate endpoint, conditional on data from a number of studies reporting the treatment effects on both outcomes.

Alternatively, a bivariate random effects meta-analysis (BRMA), in either standard form or the product normal formulation (as referred to in section 2.1.1. of this report, on multivariate meta-analysis methods) can be used if the assumption about the random effect at the between-studies level is reasonable (assuming that the true effects are exchangeable, i.e. follow a common bivariate normal distribution). In this case, predictions could be obtained with increased precision compared to when using the standard model. Daniels and Hughes, however, elaborated that the choice of the distribution for the random effects may be difficult or complex when, for example, the distribution of the effects on the surrogate endpoint is bimodal (for example when two classes of treatments of very different effectiveness, both against the same control, are investigated). Bujkiewicz et al. showed in a simulation study that when the exchangeability assumption does not hold, the predictions obtained from BRMA product normal model may be biased (Bujkiewicz et al., 2017).

An alternative approach to the choice of the between-study model could be the use of the bivariate NMA model (Bujkiewicz et al., 2019b), which is discussed in the below section 2.3.4. In the context of the distribution for the random effects, the model relaxes the assumption of normality across all trials (assumes normality only across those trials of the same treatment contrast) and as such may lead to precise predictions whilst making reasonable (or less strong) assumption about the random effects compared to the pairwise BRMA.

4.1.2 Prediction of treatment effect on the final clinical outcome given a synthesis and an effect on surrogate endpoint

Key references: (Daniels and Hughes, 1997, Bujkiewicz et al., 2019a)

Prior to making prediction of a treatment effect on the final outcome in a new study, an analysis can be carried out to investigate the predictive value of the surrogate endpoint. A cross-validation procedure (TSD 20, section 3.6.1 follows the procedure described by (Daniels and Hughes, 1997)) can be carried out to evaluate how well a surrogate endpoint predicts clinical benefit. At the same time, this process serves another purpose of assessing the model fit.

The same model can then be used to make prediction of the treatment effect on the final outcome in a new study from the treatment effect measured on surrogate endpoint in this study. Such predicted effect can then be used in cost-effectiveness model. Tan et al illustrate this conceptually in a different setting where an unreported effect on PFS and predicted from the treatment effect on OS allowing for a more detailed modelling of the natural history for metastatic prostate cancer patients treated with docetaxel (Tan et al., 2018). This is a less usual approach, as typically we would expect to have an estimate of the effect on a surrogate, such as PFS, and want to predict the effect of the treatment on OS. The paper by Tan et al illustrates the use of the methods more broadly, rather than in a typical surrogacy setting.

4.1.3 Novel methodologies for surrogate endpoints

Key references: (Bujkiewicz et al., 2019b, Bujkiewicz et al., 2016, Papanikos et al., 2020)

4.1.3.1 Multiple surrogate endpoints

Bujkiewicz et al developed multivariate meta-analytic methods in product normal formulation (see also section2.1.2 of this report) for evaluation of more than one surrogate relationship at the same time (Bujkiewicz et al., 2016). The models (which can be adapted to assume different correlation structures between treatment effects on all outcomes: the surrogate endpoints and the final outcome) can also be used for making predictions. Use of the methods can potentially lead to predicted effects obtained with reduced uncertainty as in an example in multiple sclerosis used as an illustration of the method. But this is not guaranteed as a number of other factors (such as the strength of the surrogate relationships between pairs of outcomes or the presence of treatment switching and resulting uncertainty around the effect on the final outcome) will influence both the surrogate relationship and the predictions. For example, using PFS and tumour response as combined surrogates to OS in advanced colorectal cancer did not improve the surrogacy compared to using only PFS as surrogate to OS (Elia et al., 2020).

4.1.3.2 *Surrogate relationship and mechanism of action*

Surrogate relationship depends on the mechanism of action of treatments or treatment classes. When this is the case, surrogate relationship may be investigated in subgroups. Data included in such analysis may be limited to a small number of studies per treatment class. This may dramatically reduce evidence base for surrogate endpoint evaluation. To overcome this limitation, new methods have recently been developed.

Bivariate network meta-analytic method for surrogate endpoint evaluation allows for modelling surrogate relationships in each treatment contrast individually whilst borrowing information from other treatment contrasts by taking into account the network structure of the data (Bujkiewicz et al., 2019b). The authors also proposed an extension of the method that in addition to modelling the study-level surrogate relationship (within each treatment contrast), a treatment-level surrogacy is also modelled by assuming additional similarity between the treatments. This extended method allows for predicting treatment effect on the final outcome for a new study and a new treatment, however the assumption of exchangeability is strong and needs to be considered carefully. The limitation of the bivariate NMA method may be the assumption of consistency in the network, which may not be satisfied when therapies of different mechanisms of actions are investigated in different patient populations.

Another recently developed hierarchical meta-analysis method allows for borrowing of information about surrogate relationships between treatment classes (Papanikos et al., 2020). It is a pairwise method, which does not make such strong assumption (about consistency) as the bivariate NMA method. Two versions of the method are proposed by the authors, one assuming exchangeability (similarity) of the surrogate relationships across the treatment classes and a model which relaxes this assumption by allowing for partial exchangeability, i.e. the level of exchangeability is defined by a probability of similarity which is learned from the data.

In absence of evidence from the same treatment class, appropriate meta-analytic methods (such as bivariate meta-analytic methods and hierarchical meta-analysis) should be used to validate and quantify the surrogacy relationship. This is particularly relevant for first-in-class treatments.

4.1.4 *Individual participant data (IPD) based methods for surrogate endpoint evaluation*

TSD 20 is focused on meta-analytic methods for summary data which, when applied to surrogate endpoint evaluation, describe study level surrogacy patterns, which is most relevant to regulatory decision making. When IPD are available, surrogate endpoints can be evaluated at both individual- and study-level (Burzykowski, 2006). Methods by Buyse et al. model surrogate endpoints in a hierarchical mixed model, which is de facto a meta-analytic framework (Buyse et al., 2000). The authors developed a mixed model framework where two measures, the individual-level R-squared and the trial-level R-squared, were developed to validate candidate surrogate endpoint at both levels simultaneously. The trial level part of their models for surrogate relationship is very similar to the bivariate random effects meta-analysis approach and can be used to predict the relative treatment effect on the final clinical outcome from the effect measured on the surrogate endpoint. These methods are developed in a frequentist approach. Various extensions to this approach have been proposed (Burzykowski, 2006), for example for the time-to-event data (Burzykowski et al., 2001) which used a copula function to account for the association between PFS and OS. It is a two-stage meta-analytic approach, which means that the within-study variability is modelled in a greater detail compared to any aggregate data approach (using a joint survival function), but at the study level the model is equivalent to BRMA.

4.1.5 Application and realisation of clinical outcomes on CEA and recommendations

When using a surrogate endpoint in cost effectiveness analysis, such endpoint needs to be properly validated and associated uncertainty taken into account. Ciani et al describe a three-step process to validate and use surrogate-based evidence for use in health care decision making (Ciani et al., 2017). The steps aim to (i) establish the level of evidence to consider the suitability of surrogates, (ii) assess of the strength of association, and (iii) quantify the relationship between the surrogate and the final outcome (and between the observed effect on surrogate endpoint and the expected effect on the final outcome, which will in turn impact on quality adjusted life years (QALYs) in cost-effectiveness analysis). The meta-analytic methods for surrogate endpoint evaluation, listed here and described in more detail in TSD 20, model treatment effects on a surrogate endpoint (or multiple surrogate endpoints) and the final outcome simultaneously, and as such can be used to carry out the analyses for the final two steps.

4.1.5.1 Validation

When evaluating the strength of surrogate relationship between the treatment effects on the surrogate and final outcomes, the meta-analytic methods allow for appropriate representation of uncertainty around all parameters. This is in contrast to, for example, meta-regression which ignores the uncertainty around the treatment effect on the surrogate endpoint, resulting in potentially overly optimistic estimates of surrogacy criteria and predictions (Bujkiewicz et al., 2017). Results of the analysis can be used, for example, to justify use of a surrogate endpoint as a primary outcome measure.

4.1.5.2 Making predictions

The methods described above can be used to predict an unknown treatment effect on the final outcome from the treatment effect measured on the surrogate endpoint (conditional on data from a number of studies reporting treatment effects on both outcomes). Such predicted effect, for example a hazard ratio (HR) for OS can be used in cost-effectiveness analysis.

There may be more than one new study reporting the treatment effects on the surrogate endpoint but not on the final outcome. We can then predict the treatment effects on the final outcome for all of those new studies and either treat them as separate predicted estimates that can be used in a decision-making framework individually or jointly by obtaining an average predicted effect by using, for example, a standard (univariate) meta-analysis of the predicted effects. When the new studies are investigating different treatments, individual predicted effects for each study are likely to be of interest.

4.1.5.3 Joint modelling of treatment effects on two outcomes

The models can be used to jointly model the treatment effects on both surrogate and final outcome resulting in a joint posterior distribution of relative effects on both outcomes used to inform health economic model (see also sections 2.1.5 and 2.1.7 on use of multivariate meta-analysis).

The IPD based methods allow for the evaluation of surrogate endpoints at the individual level in addition to trial level surrogacy. Knowledge about the strength of the individual level surrogacy may be helpful when modelling the natural history of a disease. However, the authors have mostly focussed on the use of these methods for validation of surrogate endpoints at both individual and trial level as well as for prediction of effect of treatment on the final clinical outcome given its observed effect on the surrogate endpoint (Buyse et al., 2000).

A limitation of the methods is that IPD from all trials in the meta-analysis are required to fit the model.

4.1.6 Examples of use of surrogate endpoints in the context of NICE

Taylor and Elston carried out a survey of NICE technology appraisals (TAs) and identified four TAs that were carried out using surrogate endpoints (Taylor and Elston, 2009). However, far more evaluations based on surrogates have been done since the review. An updated review of past TAs is needed to fully understand how surrogate endpoints have been used in decision making by NICE and to what extent the approaches have been robust and successful. In some examples of TAs, surrogate relationships have been used to predict the relative treatment effect on the final clinical outcome, and in others to extrapolate survival curve when data on OS were not sufficiently mature for use in health-economic model, such as a Markov model or a partitioned survival model. Below we list two examples for illustration.

A NICE technology appraisal committee has approved venetoclax in combination with rituximab for treating relapsed or refractory chronic lymphocytic leukaemia using PFS as a primary outcome (NICE TA561 https://www.nice.org.uk/Guidance/TA561). This was made on the basis that treatment effect on PFS was deemed a surrogate for the effects on OS. In the natural history model, survival curve for OS was difficult to obtain due to lack of maturity of the data. A joint survival model for both endpoints, PFS and OS, assuming proportionality and the same parametric form between them was used for purpose of extrapolation. However, it is unclear whether this was appropriate (see recommendations for research at the end of this chapter).

In the 2006 review of TA132, evaluating ezetimibe for the treatment of hypercholesterolaemia (https://www.nice.org.uk/guidance/ta132), the company used the effect of ezetimibe on a surrogate end point (lowering low-density lipoprotein cholesterol (LDL-c level) in the absence of clinical outcome data on the cardiovascular benefit. A previous model, developed based on trials for statins, for surrogate relationship between the LDL-c and cardiovascular effects has been used to map the effect of ezetimibe on LDL-c onto the cardiovascular benefit. The model appears to be a meta-regression, which has a limitation as it does not take into account the uncertainty around the treatment effects on the surrogate, however, the details of the method have not been listed. The ERG had a number of reservations regarding how well the surrogacy translates onto ezetimibe which is not a statin, given that the surrogacy model was developed based on trials for statins, and also the effect on surrogate was short term and thus it was unclear if it would sustain over long time horizons. The ERG recommended that ezetimibe is reassessed when more mature data are available. Ezetimibe monotherapy was recommended by the NICE TA committee as an option for the treatment of adults with primary hypercholesterolaemia who would otherwise be initiated on statin therapy but who are unable to do so because of contraindications to initial statin therapy. The committee also recommended collection of further data to establish the long-term effectiveness of ezetimibe and whether there are any long-term adverse effects.

4.1.7 Recommendations

- Meta-analytic modelling techniques described above (2.3.2 – 2.3.5) are recommended (with appropriate consideration of available data and model assumptions) when using surrogate endpoint evaluation. The assumptions are briefly discussed above. Specific details about the methods in terms data requirements, assumptions and limitations are included in TSD 20.
- Previously published models, if available, can be used to assess validity of a surrogate endpoint or to make predictions. However, they can be used only when the models have been developed in the same or relevant setting (patient population, a broad range of treatment options that suggest the surrogate relationship holds across all treatments or within a class of treatment relevant to the new technology under assessment) and are reported in sufficient detail to allow for incorporating all relevant uncertainty when making prediction. This may be difficult when a new surrogate endpoint is investigated or a new type of treatment or line of therapy is under consideration (or a new population) and surrogacy in the new context may not hold. When historical models are based on data collected in a different setting, development of a new model, using appropriate meta-analytic techniques, is recommended. This may include network meta-analysis or hierarchical methods reflecting differences in mechanism of action between treatment classes or for first-in-class scenarios.
- When evidence on the suitability of the surrogate endpoint is limited, a recommendation can potentially be made conditional on the cost-effectiveness results being confirmed when more mature data become available and the technology under investigation is re-evaluated. This is particularly recommended for cancer therapies. This approach is recommended for the licensing decisions

made by EMA based on a surrogate marker and can be adopted in HTA decision-making for cancer therapies. Coordination between NICE and EMA may improve the process.

4.1.8 Research recommendations

- Whilst the evidence synthesis methods described in this review serve well the purpose of surrogate endpoint validation and making predictions of the relative treatment effect on the final clinical outcome (which can be used in decision modelling frameworks), they do not capture all aspects of the role a surrogate endpoint can play in the decision modelling. When data on the final outcome are not mature, extrapolation of the survival curve may need to be carried out (as in the aforementioned example in leukaemia, sec 2.3.7). How such extrapolation may be carried out and what data (from RCTs, perhaps based on a surrogate endpoint, or from other sources, such as observational data) are used will depend on data availability and maturity. Optimal approaches for predicting baseline survival patterns in a natural history model in relation to use of surrogate endpoint need to be investigated. The necessary research is planned in a current NIHR fellowship application (currently the feasible timeline is 3 years).
- An up-to-date review of past technology appraisals is needed to fully understand how surrogate endpoints have been used in decision making by NICE and further methodological considerations, including simulation studies, need to be carried out to assess these methods and identify most optimal approaches in different disease and data scenarios. As above, the necessary research is planned in a current fellowship application with feasible timeline of 3 years.

5 USE OF INFORMATIVE PRIOR DISTRIBUTIONS

5.1 Informative prior distributions for between-study heterogeneity

Key references: Dias et al. (2018), Turner et al. (2019), Ren et al. (2018)

5.1.1 When the methods are likely to be useful

When meta-analyses only have a small number of studies, it is challenging to estimate the between-studies heterogeneity parameter. Similarly, for network meta-analysis, the heterogeneity can be difficult to estimate when the number of studies *per comparison* is small. At least four or five studies has been suggested as a minimum per comparison to estimate between study standard deviation (Gelman, 2006), although in a NMA with many loops (i.e. multiple sources of evidence on a single comparison) it may be possible to estimate this well with fewer studies per comparison.

When estimating random effects (network) meta-analysis models in a Bayesian context, a prior distribution must be given to the between-study heterogeneity parameter, but this can be relatively influential on the posterior distribution when there are only a small number of studies. Typically, as recommended in TSD2, minimally informative Uniform prior distributions are used, and sensitivity analyses using alternative distributions are recommended. However, poor estimation of the between-study heterogeneity due to insufficient studies can lead to implausibly wide 95% credible intervals for the relative treatment effects when typical minimally informative prior distributions are used.

However, the Bayesian approach gives the opportunity to use external information to help estimate the between-study heterogeneity. Note however that the main impact of doing this is to improve the precision of the estimates, rather than change the point estimates for the relative treatment effects.

5.1.2 Empirically based prior distributions

Turner et al. (Turner et al., 2012, Turner et al., 2015b) estimated the between-study heterogeneity for 14,886 meta-analyses in the Cochrane library with binary outcomes and obtained a prediction for what might be expected to be seen in a new study exchangeable with those Cochrane meta-analyses. They give predictions for the between-study **variance** in 80 different settings, according to the type of interventions compared, outcomes measured, and disease area. The distribution that best matches the current study can be used as a prior, when the outcome is binary and analysed on the log-odds ratio scale.

Similarly 6,492 meta-analyses in the Cochrane library with a continuous outcome were used to obtain predictive distributions for meta-analyses with different continuous outcome types, intervention comparisons type and medical area. Predictive t-distributions for the **log of the between-study variance** (on the standardised mean difference scale) expected in future meta-analyses with a continuous outcome were obtained, which can be used as informative prior distributions, with suitable re-scaling (Rhodes et al., 2015, Ren et al., 2018).

Empirically based prior distributions are not currently available for other outcome types (eg outcomes measured on the log-hazard ratio scale) although they may become available in the future.

5.1.2.1 Truncated prior distributions

It has been suggested that the empirically-based prior distributions proposed by Turner et al. (2015b) and Rhodes et al. (2016) still allow for very large, and implausible, values of the between-study heterogeneity in many cases. An elicitation method, combining empirical evidence and expert beliefs on the ‘‘range’’ of treatment effects was proposed, to infer a more informative prior distribution for the between-study heterogeneity (Ren et al., 2018). This allows a truncation of the prior distributions suggested by Turner et al. (2015b) and Rhodes et al. (2016) which is simple to implement in the standard Bayesian meta-analysis code presented in TSD2 (Ren et al., 2018, Dias et al., 2018). WinBUGS code to implement these prior distributions is available in Turner et al. (2015b), Rhodes et al. (2016) and Ren et al., 2018). An example with code is also given in(Dias et al., 2018).

5.1.3 External data to inform prior distributions

If there are insufficient data in the meta-analysis, it may be reasonable to use the posterior distribution, or a posterior predictive distribution (Lunn et al., 2013) from a larger meta-analysis on the same trial outcome involving a similar treatment for a similar condition (Higgins and Whitehead, 1996). Such an analysis could be used to approximate an informative prior distribution.

If there are no data on similar treatments and outcomes that can be used, an informative prior distribution can be elicited from a clinician who knows the field (Ren et al., 2018, Dias et al., 2018). The expert would be asked to comment on the level of variability expected in the studies, not on the actual size of the relative treatment effects. Although this would be subject to the experts’ own beliefs, it would only impact the variability in relative treatment effects, not their mean values.

5.1.4 Choice of appropriate prior distribution in NMA

In NMA, it is usually assumed that between-study heterogeneity is equal across treatment comparisons. However, the use of empirically based prior distributions (section 2.4.2) requires definition of treatment comparison type as pharmacological intervention versus control/placebo; pharmacological intervention versus pharmacological intervention; or comparison of any non-pharmacological intervention (Turner et al., 2015b, Rhodes et al., 2015). In a NMA it is likely that multiple types of comparison are included. Choice of which exact prior distribution to select is then not obvious, but the widest candidate distribution could be chosen, or the one with the majority of studies.

Models which allow for different heterogeneity parameters across comparisons can also be fitted (Lu and Ades, 2009) although it is unclear whether the added model complexity is justified. In this case, it is even more important to ensure that there is sufficient information to estimate the between-study heterogeneity across all comparisons in the network. Turner et al. (2019) proposed models for incorporating external information as prior distributions for the different variance parameters. In such cases empirically based prior distributions appropriate for each comparison type could be used to inform the different heterogeneity parameters.

5.1.5 Recommendations

- When there are few included studies, it may be preferable to use informative prior distributions, rather than the typically recommended vague prior distributions, for the heterogeneity parameter. Distributions tailored to particular outcomes and disease areas, based on studies of many hundreds of meta-analyses (Turner et al., 2012, Rhodes et al., 2015) are recommended. Truncation of these distributions may be reasonable to improve precision of treatment effect estimates (Ren et al., 2018).
- If a large meta-analysis of other treatments for the same condition and using the same outcome measures can be identified, the posterior distribution for the between-trial heterogeneity from this meta-analysis can also be used to inform the current analysis (Dakin et al., 2010).
- It is important to note the source of the prior distribution for the between-study heterogeneity and justify it use, including sensitivity of the main synthesis results to different candidate prior distributions.
- Whatever prior distribution is used for the between-study heterogeneity, results should be checked to ensure that the prior distribution was not overly informative (i.e. that it overly constrained the posterior distribution) and that it has been updated in light of the data. This can be done in by plotting the posterior density of heterogeneity parameter and comparing it to the prior distribution.

- Care should be taken to ensure the empirically-based prior distributions are implemented on the correct scale for the heterogeneity parameter: standard deviation, variance or log-variance (Rhodes et al., 2015, Turner et al., 2015b, Dias et al., 2018).
- Informative prior distributions for the relative treatment effects are not recommended unless under very specific circumstances (e.g. very sparse adverse event data) and would require additional justification.

5.1.6 Research recommendation:

- Empirically based prior distributions are beginning to be used in TA submissions, although at least one case of incorrect coding has been seen [*company submission available to York ERG, not in public domain*]. A practical guide could be produced going over how to use the empirically based prior distributions and updating the TSD2 WinBUGS code to include the Turner et al. (2015b) and Rhodes et al. (2016) prior distributions. This could be an addendum or update to TSD2 or a standalone "*How to…*" document citing TSD2. In addition, an illustration of the impact of using empirically based prior distributions on cost-effectiveness results and decision uncertainty could be included.
  Duration of project: 4 weeks
  Timeline: by September 2020

5.2 Informative prior distributions for correlation parameters

Key references: Bujkiewicz et al. (2019a) TSD20

In multivariate random effects meta-analysis borrowing of strength across outcomes is greater when correlation between outcomes is highest. Informative prior distributions on the between-studies correlation can improve borrowing of information. Alternatively, weakly informative prior distribution(s) on the between-studies correlation(s), can achieve increased borrowing, compared to the seemingly non-informative flat prior distributions on the correlation (such as uniform prior ranging between values of -1 and +1) which can result in biased pooled estimates (Burke et al., 2018).

This is described in more detail in Section 2.

6 SYNTHESIS OF SURVIVAL DATA

Studies that report survival outcomes, or more generally time-to-event outcomes, raise particular challenges for evidence synthesis. Data is collected over time, typically summarised by a Kaplan Meier survival curve (although some studies may only report median times or hazard ratios), and the shape of survival curves may differ across studies and even across treatment arms. The relative effects of treatments are measured with hazard ratios, which may either be constant over time (proportional hazards) or change over time. The validity of the proportional hazard assumption may vary with treatment comparison and study. Sections 5.1 and 5.2 describe methods for synthesis when the proportional hazards assumption does and does not hold respectively. To properly explore the proportional hazards assumption and fit models that allow for non-proportional hazards requires individual patient data (IPD), which in NICE TAs is usually available for the manufacturer's trial, but not for other trials. Section 5.3 describes methods available to approximate IPD from published Kaplan-Meier curves.

It is common (especially in oncology) for studies to report data on progression free survival (PFS) and overall survival (OS), but not on post-progression survival (PPS). OS is the sum of PFS and post-progression survival (PPS), and so OS is not independent of PFS. Sections 5.4 describes methods for joint synthesis of PFS and OS data, when IPD are available and when they are not. A particular issue that may lead to biased estimates of PPS (and hence OS) is when patients on the control arm switch to the active new treatment post progression. We discuss methods to account for such "treatment switching" in section 5.5.

Evidence from RCTs is further limited by the follow-up period of the trials. The clinical effectiveness estimates can therefore only be interpreted within the follow-up periods that have been observed. However, for cost-effectiveness, an estimate of mean survival is needed which requires extrapolation of the survival curves beyond the trial follow-up periods. We discuss methods to extrapolate pooled survival curves in section 5.6.

6.1 Synthesis of Survival Data under the Proportional Hazard Assumption

Key references: (Dias et al., 2018) (Chapter 10), (Dias et al., 2011b) (TSD2)

A review of cost-effectiveness analyses in NICE TAs found that the proportional hazards assumption is rarely checked, but commonly assumed (Guyot et al., 2011). More broadly, Krishan et al. reviewed meta-analyses of time-to-event outcomes, and found it was checked in none of the 35 meta-analyses with aggregate data, only 3/7 of the meta-analyses with a mixture of IPD and aggregate data, and only 30/81 of the meta-analyses with IPD on all studies(Krishan et al., 2019). Methods to assess the proportional hazard assumption include inspection of log-cumulative hazard plots, inclusion of time-varying covariates, inspection of Schoenfeld residuals, and visual inspection of the Kaplan-Meier curves (for example observing survival curves that cross)(Collett, 2003). Most of these methods require IPD, but may have been applied and reported in the original study publications. As long as a Kaplan-Meier curve is published, then an approximation to the IPD can be reconstructed (see section 5.3) and the PH assumption checked (although covariate information is not available in reconstructed IPD).

If the proportional hazards assumption is deemed to hold across all studies, then hazard ratios can be pooled using standard code from Evidence Synthesis TSD2 section 3.5 ((Dias et al., 2011b)), with particular care taken to properly account for correlations between relative treatment effects from RCTs with 3 or more arms (and hence 2 or more hazard ratios)(Franchini et al., 2012). Hazard ratios can be obtained from each study by fitting a Cox proportional hazards model which has the advantage that it does not assume a parametric form for the survival curve shape in each study(Collett, 2003).

If the proportional hazards assumption does not hold across all studies, then clinical input should be sought, to advise as to whether there is a plausible reason for proportional hazards for some treatment comparisons and not others. If most of the studies indicate proportional hazards holds and there is no clinical explanation why others would not, then it may be justified to pool hazard ratios, although sensitivity analysis to this choice is advisable using the methods presented in section 5.2.

If Kaplan-Meier curves are not available, and different studies report different summary measures (for example medians, or proportions surviving to specific time-points, or HRs may have been reported in different studies), then as long as a parametric functional form is assumed, these different summaries can be combined in a pairwise or network meta-analysis using a shared parameter model. For example, if an Exponential model is assumed, then each outcome type provides information on the event rate: the median provides evidence on 2/rate, the proportion surviving to time t provide information on exp(-rate*t), and the HR provides information on the ratio of the rates. If two-parameter models are assumed (eg Weibull) then it is necessary to assume something about one of the parameters in order to estimate the other (e.g. a constant or exchangeable shape parameter across studies(Welton et al., 2008, Welton et al., 2010).

Saramago et al. present a shared parameter method to combine studies with IPD time-to-event data and aggregate count data to estimate a Weibull survival model, including covariate interactions(Saramago et al., 2014).

6.1.1 Recommendations

- The proportional hazards assumption should always be assessed, preferably using log-cumulative hazard plots (as advised in TSD14 ((Latimer, 2011))), visual inspection of the Kaplan-Meier curves, and interpretation of tests for proportional hazards reported in the original trial publications.
- If the proportional hazards assumption holds then hazard ratios may be pooled using standard code for treatment differences (Evidence Synthesis TSD2 section 3.5(Dias et al., 2011b)). Correlations need to be accounted for in trials with 3 or more arms.
- Shared parameter models may be possible when different studies report different summary outcome measures by assuming a common underlying survival model. However, reconstructing Kaplan Meier curves to obtain common summaries across studies is preferable if possible.

6.2 Synthesis of Survival Data under non-Proportional Hazards

If the proportional hazards assumption does not hold in some or most of the studies, then it is not appropriate to synthesise the hazard ratios, because they depend on time and cannot be described by a single summary. A variety of alternative methods have been proposed in the literature for this situation. We first briefly recap methods for analysis of survival data from a single trial and then discuss how these methods extend to synthesis of multiple trials and treatments.

6.2.1 Modelling non-proportional hazards in the analysis of a single trial

Parametric survival models may be fitted if IPD (or reconstructed IPD) are available, and these may be parameterised to assume non-proportional hazards. The most intuitive alternatives to proportional hazards are accelerated failure time (AFT) models where it is assumed that the effect of treatment is to accelerate or decelerate time to event by a proportional constant (the acceleration factor)(Collett, 2003). Distributions where AFT models can be fitted include log-logistic, log-normal, gamma, and Weibull distributions. For example, the Weibull distribution can be parameterised (i) assuming proportional hazards (and estimate a hazard ratio), (ii) as an AFT model (and estimate an acceleration factor), or (iii) to reflect other departures from proportional hazards by estimating treatment effects on both shape and scale parameters (equivalent to fitting separate distributions to each arm). Methods to select an appropriate parametric form in the analysis of a single trial are given in TSD 14(Latimer, 2011), with a preference for using the same parametric form for all arms within in a trial.

More flexible survival models include piecewise models(Friedman, 1982), cubic spline models (Royston and Parmar, 2002) and fractional polynomials (Royston and Altman, 1994). These models have the advantage that they can capture a wider range of survival curve shapes than the standard parametric survival curves, although they may be more complex to fit, less easily interpreted, and can give unrealistic extrapolations.

Finally, Royston and Parmar proposed a non-parametric approach which measures the treatment effect as a difference or ratio of restricted mean survival time (RMST) across treatment arms(Royston and Parmar, 2011, Royston and Parmar, 2013). The RMST is defined as the area under the survival curve up to a restriction time-point (for example trial follow-up). The method is simple and makes few assumptions, but is only valid up to the restriction point of the trial, and does not deliver an estimated survival curve for use in modelling.

6.2.2 Synthesis of Survival Data under non-Proportional Hazards

If an AFT model is appropriate across the collection of studies/treatment comparisons, then pairwise or network meta-analysis models can be used to pool the log of the acceleration factors. Standard code from Evidence Synthesis TSD2 section 3.5 can be applied (although we are not aware that this has been applied in practise to date), with particular care taken to properly account for correlations between relative treatment effects from RCTs with 3 or more arms (and hence 2 or more hazard ratios)(Franchini et al., 2012). Note that this represents a 2-stage analysis, as the acceleration factors are estimated from each trial in a first stage, and then they are pooled in the second stage. Siannis et al. proposed a one-stage approach for meta-analysis of acceleration factors (they denoted these percentile ratios)(Siannis et al., 2010) and also a 2-stage version (Barrett et al., 2012). We are not aware that this method has been extended to NMA, although there is a conference abstract available(Bexelius et al., 2014).

If a particular parametric survival model is appropriate across all studies and treatments, then the parameters from that model can be pooled. For example, a bivariate meta-analysis model (see section 2.1.1) can be put on the shape and scale parameters of a Weibull model, where shape and scale depend on treatment and study ((Welton et al., 2008, Welton et al., 2010) (Ouwens et al., 2010)), which can be generalised to other parametric distributions (Cope et al., 2017) and Dias et al (2018) Chapter 10). The same approach can be used with piecewise models. The most common such model is the piecewise exponential model, where exponential distributions with different rates are assumed for different segments of the curves. Relative treatment effects are assumed to act on the hazard ratios for each segment, and a multivariate meta-analysis model put on the piecewise log HRs(Lu et al., 2007).

Synthesis of survival outcomes using fractional polynomial models (Jansen, 2011) have been used in several manufacturer submissions to NICE (for example TA463 Cabozantinib for previously treated advanced renal cell carcinoma (https://www.nice.org.uk/Guidance/TA463)). Fractional polynomial models can be thought of as a family of parametric curves which cover a wide range of shapes for the survival curves. A first order fractional polynomial model has 3 parameters, one of which (the power parameter) is usually fixed. A second order fractional polynomial model has 5 parameters, two of which are fixed power parameters. Although in theory all of these parameters could depend on study and treatment, it is sensible to fix power parameters (chosen using model fit and parsimony metrics), and use a multivariate meta-analysis (or NMA) model for the remaining 2 or 3 parameters, exactly as for parametric models. The advantage of using fractional polynomial models is that they can capture a wide range of survival curve shapes and they can pool studies where the shape varies between study and treatment. This is particularly advantageous when standard parametric models do not provide a good fit to the data. However, this means that different survival shapes may be estimated for the same treatment in different studies, which may not have clinical validity. The parameters of flexible polynomial models are not intuitive, and so relative effects may be difficult to interpret and “sense-check”. Finally, the flexibility of the curves may mean that they “over-fit” the observed data and may give predictions that do not have face-validity. For example in TA463 the fractional polynomial models fitted to OS and PFS crossed, so there was part of the curves where OS was less than PFS (which cannot happen).

Spline models have been used for the analysis of a single trial in NICE TAs (eg TA417 Nivolumab for previously treated advanced renal cell carcinoma (https://www.nice.org.uk/guidance/ta417)), and more recently network meta-analysis using cubic spline models have been developed to describe the log-hazard over time(Freeman and Carpenter). However, we are not aware of these yet being applied in NICE TAs. We would anticipate that these methods would have the same advantages and disadvantages as the fractional polynomial models, although they have the additional restriction that the “knots” which define the sections for the cubic splines, are placed identically across the studies (as for piecewise exponential models).

Meta-analysis of differences or ratios of RMSTs has been proposed(Wei et al., 2015), and extended to network meta-analysis by Daly et al. (unpublished but used in NICE Guideline NG122 Lung cancer: diagnosis and management (https://www.nice.org.uk/guidance/ng122)). The advantage of this approach is that it is non-parametric and so does not make any assumptions about the survival shapes in the different studies and treatment arms. Each study provides information on the difference in area between the two survival curves (or the ratio of these areas). Under this approach features such as crossing survival curves (a sure sign that proportional hazards does not hold) are dealt with naturally, the additional survival when the curve for A is higher than that for B is subtracted from the additional survival when the curve for B is higher than that for A. The main drawback of this approach in the context of NICE TAs is that the method requires that the restriction time used to calculate the RMST is the same across the studies, and all included studies must have follow-up at least up to the restriction time. This means that we are limited to make treatment comparisons for the period of the shortest trial follow-up, and any data beyond that point is not incorporated in the synthesis. Daly et al suggest combining the approach with extrapolation methods to overcome this limitation, but note extrapolation brings further assumptions (see section 5.6). Another limitation of this approach is that because the method is non-parametric no survival curve is estimated, which is required for most cost-effectiveness models. However, Daly et al show how partitioned survival models can be estimated using the results from a synthesis of differences or ratios of RMSTs, including incorporating discounting.

6.2.3 Recommendations

- If the accelerated failure time assumption is reasonable across all studies, then acceleration factors may be pooled using a 2-stage approach using standard code for treatment differences (Evidence Synthesis TSD2 section 3.5). Correlations need to be account for with trials with 3 or more arms.

- If the same parametric model is reasonable across studies and treatment arms, then multivariate pairwise or network meta-analysis on the survival curve parameters may be fitted (Cope et al. 2017, Dias et al 2018 Ch 10).

6.2.4 Research Recommendations

- There is a need for research to explore the use of pairwise and network meta-analysis of differences or ratios of RMSTs to inform cost-effectiveness models typical in NICE TAs. This work is already at an advanced stage, and two manuscripts in early draft form.
- There is a need for a TSD giving a critique of the different methods for synthesis of survival outcomes when the proportional hazards assumption does not hold.

6.3 Reconstructing Survival Data from Published Kaplan-Meier Curves

When it is not possible to synthesise a single summary (eg hazard ratio or acceleration factor) that is reported in all studies, synthesis of survival outcomes requires IPD for all included trials. Although IPD is usually available from the manufacturer's trial, this is not the case for the other trials. However Kaplan-Meier curves are usually available. There are a variety of digitising software options available to obtain points from the Kaplan-Meier curves by clicking at multiple time points of the curves, so long as the publication quality is sufficiently clear (not too blurry). Ideally these timepoints should capture the "steps" in the Kaplan-Meier curves where events occur (Guyot et al. (2012)).

Guyot et al. (2012) developed an algorithm that constructs a dataset that produces a Kaplan Meier curve that approximates the published curve. The method uses the extracted points from the Kaplan-Meier curve and other information, such as the reported numbers at risk under the survival curve, and, the total number of events and total number censored. Guyot et al. (2012) found, that the method gives a high level of accuracy for survival proportions and medians, and a reasonable degree of accuracy for HRs when numbers at risk under the curve or total number of events are reported. The method is frequently used in submissions to NICE.

The Guyot method has the advantage that it reconstructs the IPD to enable re-analysis, whereas previous methods obtain survival probabilities at a limited set of specific time points ((Parmar et al., 1998, Williamson et al., 2002)) and/or impose parametric models to obtain the probabilities ((Dear, 1994, Arends et al., 2008, Fiocco et al., 2009, Ouwens et al., 2010, Jansen, 2011, Hoyle and Henley, 2011)). The Guyot method was developed as an R routine, but has also been implemented in Stata(Wei and Royston, 2017).

Note, that reconstructing Kaplan-Meier data does not provide patient level data for covariates unless Kaplan-Meier curves are published separately for different subgroups, and even if this is the case covariates cannot be modelled jointly. This limits the use of the method for subgroup analysis.

6.3.1 Recommendations

- The Guyot (2012) method can be used to reconstruct Kaplan-Meier data. Implementations are available in R or Stata. Covariate information can only be obtained if Kaplan-Meier curves are published by sub-group.

6.4 Combined Analysis of PFS and OS

PFS and OS are typically analysed separately in HTAs, and the resulting estimates used to construct a partitioned survival cost-effectiveness model. However, PFS and OS are not independent. Firstly, OS is the sum of PFS and post-progression survival (PPS), and so will be correlated with PFS. Furthermore, time to progression may be correlated with PPS which further leads to correlations between PFS and OS. It is for these reasons that PFS is often considered a surrogate for OS (see section 2.3). The methods in section 2.3 can therefore be applied to jointly model PFS and OS, and we do not repeat those methods in this section. Here we discuss specific alternative approaches that have been proposed for joint modelling of PFS and PPS and discuss their advantages and disadvantages over the joint models presented in section 2.3.

6.4.1 When actual IPD are available

If IPD (actual IPD, not reconstructed IPD) are available from each study, then a state-transition model may be a more appropriate approach than the partitioned survival approach (Williams et al., 2017, Woods et al., 2017). For survival outcomes, state transition models consist of movements between 3 states: progression-free, post-progression, and dead, with survival models fitted for PFS and PPS which may include dependence between time to progression and PPS. Actual IPD are required (not reconstructed) because the method needs to know both time of progression and mortality for each individual (which cannot be reconstructed from PFS and OS Kaplan-Meier curves).

Methods for evidence synthesis of multi-state model parameters with IPD have rarely been conducted, because it is unusual to have actual IPD for all studies. Price et al. give methods for evidence synthesis for transitions between asthma health states (Price et al., 2011). (Jansen and Trikalinos, 2013) describe an application of network meta-analysis with fractional polynomials applied to a multi-state model for survival outcomes, but this is just a conference abstract. If Kaplan-Meier curves are available for PFS and PPS, then simple multi-state transition models that do not allow a dependence between PFS and PPS can be estimated from reconstructed IPD. However, curves for PPS are rarely reported.

Whilst state-transition models have been used in NICE TAs (see review in TSD19), this is for the analysis of the manufacturer's trial. TSD 19 highlights the lack of available methods to incorporate indirect comparisons and network meta-analysis in multi-state models. Price et al. discusses a range of modelling options for parameterising treatment effects on Markov model state transition rates(Price et al., 2011).

The bivariate meta-analysis methods for joint modelling of PFS and OS (section 2.3) only require actual IPD to be available from at least one study to estimate the relationship between treatment effects on the two outcomes. However, although they can be fitted when only reconstructed IPD are available from other studies, very strong assumptions are required on the similarity of the treatment effect relationships between PFS and OS in different studies and treatments, which may not be plausible.

6.4.2 When actual IPD are not available

Daly et al (unpublished but used in NICE Guideline NG122 Lung cancer: diagnosis and management (https://www.nice.org.uk/guidance/ng122)) present a non-parametric joint model for relative treatment effects pooled as differences or ratios of restricted mean survival time (RMST) for PFS and PPS. RMST is estimated by the area under the survival curves (AUCs) for PFS and OS, and correlations imposed by the constraint that OS>PFS are captured using non-parametric bootstrap sampling. AUCs for PFS and OS are pooled, but the network meta-analysis model for relative treatment effects is put on PFS and PPS, where OS=PFS + PPS is informed by the sum of these two NMA models. This approach has the advantage that it does not make any distributional assumptions about the shape of the survival curves across treatments and studies, and accounts for correlations induced due to the structural relationship between PFS and OS. It also has the advantage that treatment effects are put directly on PFS and PPS (in line with state transition models). However, in common with bivariate models for OS and PFS (section 2.3.2) it does not allow for a relationship between PFS and PPS (reconstructed IPD are insufficient to estimate this).

6.4.3 Recommendations

6.4.4 Research Recommendations

- Research is required to develop robust methodology for the pairwise network meta-analysis to populate state transition survival models.
- Research is required for the development of non-parametric joint models for PFS and PPS based on published PFS and OS curves (the is on-going with 2 manuscripts in early draft format).
- Research is required to compare the performance of different joint modelling approaches in example typical in NICE TAs.

6.5 Treatment Switching

Key references: (Henshall et al., 2016, Latimer et al., 2019, Latimer et al., 2020, Sullivan et al., 2020)

6.5.1 Reporting Analyses & Data Requirements

TSD16 (Latimer and Abrams, 2014) discuss commonly used methods for adjusting RCTs when treatment switching (or crossover) occurs following randomisation. The main methods (Rank Preserving Structural Failure Time Models [RPSFTM], Inverse Probability of Censoring Weights [IPCW] and Two Stage Estimation [TSE]) are now being routinely used in many NICE oncology TAs. However, each of the methods require a number of assumptions to be made and have different data requirements. Whilst Latimer and Abrams (2014) discuss situations in which extensive simulation studies have identified when some methods are unlikely to be appropriate, in many situations multiple methods can potentially be used and consequently clear and transparent reporting of analyses is crucial to allow appropriate critique (Sullivan et al., 2020). The main methods also require different data requirements and often the ability of analysts to apply particular methods can be dictated by data availability. Consequently in situations when treatment switching might be expected and Overall Survival (OS) is a key endpoint, care and thought is required to ensure that appropriate data are collected, especially during the course of the trial (to enable better prediction of which patients switch), at the point of switching (to enable comparisons of those patients who switch with those who do not) and post switching (so that post switching treatments can be assessed). Such design considerations need to be co-developed in collaboration with all stakeholders – patients/public, regulators and reimbursement agencies (Henshall et al., 2016).

6.5.2 Re-censoring

RPSFTM and TSE approaches estimate counterfactual survival times (i.e. those which would have been observed had switching not occurred). As such they involve shrinking survival times for patients who switched treatments, but some patients are censored. If switching is thought to be associated with prognosis, this means shrunken censoring times could be related to prognosis and therefore associated with bias. To reduce this effect re-censoring is often applied (Latimer and Abrams, 2014). Whilst re-censoring may avoid or reduce the bias associated with shrunken censoring times being related to prognosis, it generally involves a loss of longer term information. In circumstances where the treatment effect is not constant over time this can also lead to biased estimates of the longer term treatment effect. Latimer et al. investigated via simulation studies the potential impact of re-censoring on bias and provide guidance for those undertaking such analyses (Latimer et al., 2019).

6.5.3 Two-Stage Estimation (TSE) and Covariate Selection

TSE requires a secondary baseline to be established close to the point of potential switching. For example, in many oncology trials this may be disease progression. However, in situations when there is a significant time delay between the secondary baseline and patients actually switching there is the possibility of time-dependent effects to be introduced, i.e. confounding. Latimer et al. investigated an extension to a simple TSE approach which uses structural nested models and g-estimation to account for time-dependent confounding and evaluated it using extensive simulation studies (Latimer et al., 2020). This performed well and the bias associated with g-estimation approach to account for time-dependent confounding being lower than for simple TSE, though IPCW and RPSFTM could also perform well in such circumstances.

A number of standard methods (IPCW and TSE) require the development of two statistical models, and the issue of covariate selection in both models is a key consideration. It is also one which has received relatively little attention, and has clear parallels with other areas (e.g.

joint modelling (Rizopoulos, 2012)) but requires both data availability (see above on appropriate data collection strategies) and an appropriate understanding of the causal disease-treatment pathway (see below on Causal Inference).

6.5.4 Using External and Aggregate Data & Bayesian Methods

To date there have been few TAs which have used external data to deal with the problem of treatment switching (TA171 https://www.nice.org.uk/guidance/ta17 and TA269 https://www.nice.org.uk/Guidance/TA269). However, with the increasing availability of IPD from RCTs and from Real World Evidence (RWE) studies, such an approach is likely to become more attractive and indeed feasible. However, not only are the a number of approaches which could be adopted, for example matching, calibration and prediction/imputation (Ishak et al., 2011) many of these rely crucially upon the development of a statistical model to adjust for potential patient differences, thus introducing additional model/structural uncertainty – which covariates should be used and does model choice matter. Further research into the potential uses of such methods appears to be both warranted and necessary.

All the adjustment approaches discussed so far require access to IPD. However, in a number of situations only aggregate/summary data are available but nevertheless treatment switching is known to be an issue. For example, when conducting a MAIC (See 1.2 and 1.3). In such situations if adjusted analyses have been adequately presented this may not be an issue, but unfortunately this is seldom the case. Potential solutions include adjusting aggregate/summary results based on study characteristics and knowledge of treatment switching or recreation/simulation of IPD (See 5.3) including timing of treatment switching in order that more standard adjustment methods (e.g. RPSFTM, IPCW, TSE) can then be applied before further syntheses/analyses (e.g. MAICs) are undertaken (Boucher et al., 2014).

The use of Bayesian methods for both the use of external data and when IPD are not available have been advocated and would appear attractive warranting further research. However, the use of a Bayesian approach to incorporate uncertainty regarding model assumptions (for example the constant treatment effect in the RPSFTM approach) (Sullivan et al., 2020) and use of appropriate prior distributions to model *a priori* causal effects (Oganisian and Roy, 2020) both appear to be worthy of investigation.

6.5.5 Causal Inference, Statistical Modelling & Treatment Sequences

In terms of both the choice of approach to dealing with treatment switching (and the plausibility of the assumptions that different methods make) and the development of appropriate statistical models (e.g. covariate selection) a clear understanding of the causal disease-treatment pathway is required (van Geloven et al., 2020). The use of increasingly standard causal inference approaches (e.g. causal diagrams and doubly robust methods of treatment estimation) have a potential key role to play, but these have received relatively little attention to date (Funk et al., 2011, Hernán and Robins, 2020).

As highlighted above in terms of covariate selection, there are close parallels between some of the standard adjustment methods (e.g. IPCW) and a statistical joint modelling approach (Rizopoulos, 2012). Although TAs are not concerned with estimating the effect of treatment sequences *per se* a joint modelling approach to the dynamic nature of treatments received in RCTs allowing treatment switching (Bhattacharjee, 2019) and/or the use of statistical methods to model treatment sequencing (Zheng et al., 2017, Deniz et al., 2018) could provide a fruitful area of future co-development for estimating line-specific treatment effects on OS in TAs.

6.5.6 Recommendations

- Updating of TSD16 to reflect; appropriate clear and transparent reporting of analyses which have adjusted for treatment switching, data collection strategies to enable appropriate methods to be applied, use of re-censoring for counterfactual methods, and two-stage estimation using g-estimation when there is a possibility of time-dependent confounding between secondary baseline and treatment switching.

6.5.7 Research Recommendations

- Use of external data for adjusting for treatment switching (including use of Bayesian methods)
- Use of Bayesian methods generally regarding model assumptions (and parameters – for example the common treatment effect assumption in RPSFTM) and causal effects
- Methods for the adjustment of aggregate/summary data when treatment switching is an issue
- Use of causal inference approaches to delineate the disease-treatment pathway and aid covariate selection strategies especially for two-stage statistical approaches (e.g. IPCW, TSE and use of external data)
- Use of statistical (joint) modelling of dynamic treatment strategies and treatment sequence modelling

6.6 Synthesis of RCT and External Evidence for Extrapolation

Methods for extrapolation of survival curves mostly lie outside the remit for this evidence synthesis review, however here we briefly summarise methods where evidence is synthesised to jointly estimate survival curves. We do not cover the situation where an external data source is used to select a plausible extrapolation of the manufacturer's trial data, nor where an external data source is used for the reference treatment survival curve on which hazard ratios from a pairwise or network meta-analysis are applied, although these are common uses of external data for extrapolation in NICE TAs (TSD14 (Latimer, 2011)). For a general review of use of external evidence in extrapolation see TSD14(Latimer, 2011) and Jackson et al. (Jackson et al., 2017).

Demiris and Sharples (Demiris and Sharples, 2006) use a Bayesian evidence synthesis approach to jointly estimate a survival curve on the reference treatment using a disease specific cohort data from 2 hospitals and also general population mortality data. They explored additive and multiplicative hazards models to relate mortality in the disease specific populations with the general population to extrapolate survival on the reference treatment, on which hazard ratios from a meta-analysis were applied. The advantage of this approach is that it captures effects of ageing on mortality available from general population data and calibrates this to the disease-specific populations of the representative cohorts. However, the hazard ratio from the meta-analysis is assumed to hold beyond the follow-up period of the trials.

Guyot et al. (Guyot et al., 2016) also used a Bayesian multi-parameter evidence synthesis approach to jointly estimate survival curves combining RCT data with external information on general population survival, conditional survival from the SEER cancer registry database (https://seer.cancer.gov/), and expert opinion. The method was applied to Kaplan Meier curves from TA145 on cetuximab for the treatment of locally advanced squamous cell cancer of the head and neck (https://www.nice.org.uk/guidance/ta145). The approach estimates the survival curves using the RCT data along with assumptions about the relationship between survival in the RCT population and the external data sources. They assumed that survival on the control arm is less than that in a matched general population, that 1-year conditional survival converges to that in a matched population in the SEER registry, and that the HR is a smooth function that decreases initially and then increases until it converges to 1. These assumptions were supported by a meta-analysis of a different therapy in the same population, visual inspection of 1-year conditional survival and expert opinion on mechanism of action of the drug. The method worked well on the TA145 data, but has yet to be applied to other examples where the relationships between external data and RCT survival may be different or less clear.

Both Demiris and Sharples (2006) and Guyot et al (2016) found that standard parametric curves were not flexible enough to capture the shape of the survival curves reflecting all the evidence sources. Demiris and Sharples (2006) used semi-parametric models which may be problematic to apply in cost-effectiveness models, and Guyot et al (2016) used spline models which captured the shapes of the survival curves well, but are difficult to fit (fractional polynomials or piecewise constant models may provide a more practical alternative).

There is a Technical Support Document (Rutherford, 2020) currently in preparation which will discuss flexible modelling and the need to include background mortality as an important aspect of long-term extrapolation.

6.6.1 Research Recommendation

- Research is needed to test these methods out in a range of NICE TAs to see how useful they are in practise.

7 RELIABILITY OF RECOMMENDATIONS BASED ON NMA

7.1 GRADE-NMA

K*ey references: (Puhan et al., 2014)*

The GRADE extension to NMA (Puhan et al., 2014) builds on the standard GRADE (High, Moderate, Low, Very Low) quality assessment applied to pair-wise meta-analytic summaries (Balshem et al., 2011). The GRADE-NMA extension is a set of rules that "map" from the pair-wise assessments on the "direct" evidence summaries onto a set of quality assessments on the NMA estimates. The process involves the following steps

1. Generate the standard GRADE ratings based on the direct evidence for each pair-wise contrast
2. Generate a GRADE rating for every indirect estimate (ie using indirect evidence only): it will be the lowest of the constituent ratings
3. Generate a GRADE rating for the NMA: it will be the highest of the direct and indirect ratings, UNLESS the direct and indirect estimates are substantially different.

Where direct and indirect estimates are different, the advice is to choose the estimate with the higher quality rating.

This is against every known principle of decision making. It could lead to choosing the worst treatment, and it could generate incoherent situations where A is more effective than B, B is more effective than C, and C is more effective than A.

The GRADE-NMA advice was subsequently "clarified"(Hultcrantz et al., 2017, Brignardello-Petersen et al., 2018). In a third clarification (Brignardello-Petersena R et al., 2019) the potential for incoherence was recognised, but decision makers were advised to choose the estimate with the higher quality rating, whether or not the set of estimates was now incoherent.

GRADE-NMA would be very time-consuming in large networks.

7.2 6.2 CINeMA

Key references (Salanti et al., 2014, Institute of Social and Preventive Medicine University of Berne, Nikolakopoulou et al., 2019)

CINeMA (Confidence in Network Meta-Analysis) is a web application based on (Salanti et al., 2014). It starts from the individual domain ratings on the each study with direct evidence for each pair-wise contrast, not on the summary rating. CINEMA evaluates the influence of each piece of evidence on each NMA estimate, using the “contributions matrix”(Krahn et al., 2013) which measures how much weight each study has in each pairwise summary estimate. CINEMA then formally combines these to deliver a quality rating for each NMA estimate. In comparison to GRADE-NMA, which is non-sensical, the methodology is rigorous, easier and less time-consuming to use, and cannot produce incoherent conclusions.

An important limitation of CINeMA is that it cannot be applied to the results from a Bayesian NMA. It’s input must be from a frequentist analysis, and at present it only accepts input from the *netmeta* R suite(Rücker et al., 2015).

If a quality rating is deemed useful, CINeMA is a reasonable way to generate it, subject to its inherent software limitations. However, as with GRADE-NMA, it is not clear how the individual quality ratings are to be used. How should they impact on the decision, or the decision-making process? Evidence is very often deemed to be “low” or “very low” quality: but a treatment recommendation still has to be made.

7.3 Threshold analysis

Key references (Phillippo et al., 2018b, Phillippo et al., 2019)

An alternative approach is to ask the question: “OK, maybe the evidence on A vs B is biased – but how biased would it have to be before it changed the treatment recommendation?”

This is a form of threshold analysis proposed by (Caldwell et al., 2016) with the computation of the methods developed by (Phillippo et al., 2018b), using similar methods to CINeMA. Threshold analysis can be applied to examine the influence of individual trials, of summary estimates of pair-wise contrasts, or of subsets of trials – for example all trials on a class of treatments, or all trials carried out in a particular setting. It can also be used to look at the impact of evidence regarded as “inconsistent” with the rest of the evidence network, or to any estimate considered at risk of bias

Threshold analysis allows decision makers to focus attention just on those pieces of evidence to which the decision is sensitive. Threshold analysis has recently been applied in NICE Guideline development (Inducing of Labour (Guideline in development GID-NG10082), and Specialist Neonatal Respiratory Care in pre-term babies (Guideline in development GID-QS10137)). Experience so far suggests that guideline developers use threshold analysis to directly address and potentially mitigate stakeholder criticism of particular pieces of evidence; or to moderate recommendations, for example recommending that any of a set of active treatments are chosen, rather than identifying one as the “best”.

The method is designed to take joint posterior summaries from MCMC simulation as its input, and so is compatible with any degree of complexity in the evidence synthesis, including the synthesis of multiple endpoints. (GRADE-NMA and CINeMA would have to treat different outcomes separately). However, output from frequentist analyses, in the form of mean relative effects and their covariance matrix, can be analysed as well. An R package is available: https://cran.r-project.org/package=nmathresh

The method can be applied not only to NMA outputs, but also to outputs from cost-effectiveness analysis, so that the threshold analysis is related to recommendations based on efficacy – and on cost-effectiveness. It is however limited to linear, or linearizable, decision models. Further development is underway with MRC funding to allow the method to be applied to non-linear models, incorporating for example Markov models, and – potentially -individual patient simulation models.

7.4 Recommendations

- CINeMA could be used if a confidence rating on one or more NMA estimates is required.

- Threshold analysis should be used to assess the impact on guidance of any potentially controversial aspects of the data. It could also be considered for routine use. Once the work on extending it to highly non-linear models is complete (about 18m from now), a TSD describing its use should be prepared.

8 METHODS FOR META-ANALYSIS OF COMPARATIVE ACCURACY OF DIAGNOSTIC TESTS

8.1 Introduction

Although our focus is on test comparisons, this document must be viewed in the context of guidance on meta-analysis of diagnostic test accuracy more generally. Current sources of guidance include the NICE Diagnostic Assessment Programme (DAP) manual (NICE, 2011), the Cochrane Handbook for Diagnostic Test Accuracy Chapter 10 (Macaskill et al., 2010) and the Methods Guide to Medical Test Reviews of the Agency for Healthcare Research and Quality (Chang et al., 2012). These documents focus primarily on synthesis of sensitivity and specificity evidence on a single test, and appropriately recommend bivariate random effects meta-analysis (Reitsma et al., 2005, Chu and Cole, 2006) or the hierarchical summary receiver operating characteristic (HSROC) model (Rutter and Gatsonis, 2001), which are equivalent when no covariates are fitted.

However, further guidance on the single test case is required (see Recommendations), as a necessary preliminary exercise, before proceeding to the question of test comparisons. In particular, guidance is needed on methods that accommodate estimates of sensitivity and specificity at more than one diagnostic threshold per study (Steinhauser et al., 2016, Jones et al., 2019), and how to determine the optimal threshold, which is a function of disease prevalence and net benefits attaching to true and false negatives and positives, as well as sensitivity and specificity.

8.2 Methods for meta-analysis of comparative accuracy

Key references: (Trikalinos et al., 2014b, Menten and Lesaffre, 2015)

To minimise bias, comparative accuracy is best estimated from comparative primary studies, e.g. paired studies in which each individual receives both index tests plus the reference standard (Takwoingi et al., 2013). There is no clear guidance yet, however, on how best to analyse data from such studies.

8.2.1 <u>Test as a covariate</u>

A simple approach is to include test as a covariate in the bivariate or HSROC model (Macaskill et al., 2010). This, however, ignores both within and between-study correlations arising from studies evaluating more than one test. The impact of ignoring such correlations in this context does not appear to have been investigated in depth. However, based on the more general multivariate meta-analysis literature(Riley et al., 2007, Trikalinos et al., 2014a), we would anticipate: (i) summary estimates of sensitivity and specificity, and of comparative accuracy measures (e.g. differences in sensitivity and specificity) to be relatively robust to this; (ii) but potentially a substantial impact on the *precision* of any kind of comparative measure. Notably, ignoring correlations will impact upon the level of decision-uncertainty in an economic model, although the likely extent of this impact is currently unknown. See also 7.2.2 regarding estimation of parameters representing the 'joint' accuracy of tests used in combination.

8.2.2 <u>Data inputs and model outputs</u>

Primary test accuracy studies with a paired design might report either 'fully cross classified data' (which includes, for example, the number of individuals with the disease who were positive on Test A but negative on Test B) or only 'marginal counts' (where the *overlap* between Test A and B results in the diseased and disease-free populations is unknown). Many authors have noted that fully cross classified data are rarely available from published study reports (Macaskill et al., 2010, Hoyer and Kuss, 2018b). In other words, primary studies often only report the number of true positive, false negative, true negative and false positive results (a '2 x 2 table') for each test separately, rather than reporting the full overlap (e.g. a 2 x 2 x 2 table). For this reason, most methodological development has focused around modelling of marginal counts, i.e. data from 2x2 tables (Nyaga et al., 2018a, Nyaga et al., 2018b, Owen et al., 2018, Hoyer and Kuss, 2018a, Hoyer and Kuss, 2018b), whereas only a few proposed models require fully cross classified data (possibly imputed in some studies) (Trikalinos et al., 2014b, Dimou et al., 2016, Cheng, 2016, Menten and Lesaffre, 2015).

The key limitation of modelling only marginal counts is that within-study correlations (arising when two or more tests are evaluated on the same individuals) cannot then be accounted for. Each of the models listed above, however, does account for the other source of correlation that is expected in these data: that arising between-studies. As with the situation of ignoring both sources of correlation (see 7.2.1), ignoring within-study correlations can be expected to impact upon the precision of any comparative accuracy measure. More specifically, as these correlations can generally be expected to be positive, comparative measures can be expected to be unduly imprecise, although, again, the likely extent of this additional imprecision in practice is unclear (Macaskill et al., 2010, Trikalinos et al., 2014b). Note that if comparative primary studies used a non-paired design (e.g. individuals were randomised to Test A or Test B without crossover) then within-study correlations are not present, such that the above does not apply.

We note that questions asked by the DAP often require an understanding of the *joint* accuracy of two or more tests used in combination or in sequence (Trikalinos et al., 2014b, Cheng, 2016, Novielli et al., 2013). To quantify this, knowledge and modelling of within-study correlation is crucial: estimates of joint accuracy may be extremely biased if this is ignored.

An appealing feature of the Trikalinos et al. (2014b) model is the explicit estimation of parameters representing joint accuracy. However, the model is of high dimensionality and computationally demanding, even for the case of two index tests versus a gold standard. In our experience, the model can run in to computational difficulties even for seemingly simplistic data sets.

8.2.3 Parameterisation

Most proposed models for comparative accuracy have the bivariate model parameterisation (of Reitsma et al., 2005, Chu and Cole, 2006) at their core, whereas the model proposed by Lian et al. (2019) instead extends the HSROC parameterisation. Further, most models are 'arm-based' (i.e. parameterised in terms of the absolute sensitivity and specificity of each test), whereas 'contrast-based' parameterisations are also an option (Menten and Lesaffre, 2015). These models also vary in their approach to allowing for between-study dependencies, including use of multivariate normal distributions (Hoyer and Kuss, 2018a, Ma et al., 2018, Trikalinos et al., 2014b, Dimou et al., 2016), 'analysis of covariance' (ANOVA) type formulations (Owen et al., 2018, Nyaga et al., 2018a) or use of copulas (Hoyer and Kuss, 2018b, Nyaga et al., 2018b). Further research is needed before we are able to recommend specific parameterisations (see 7.3.4).

8.2.4 Network meta-analysis of test accuracy

Most authors have focused on the case of data being available on just two tests versus a gold standard (Trikalinos et al., 2014b, Dimou et al., 2016, Hoyer and Kuss, 2018b, Hoyer and Kuss, 2018a). However, a small number of models have been proposed for networks of evidence, where, e.g. Study 1 reports data on tests A,B,C; Study 2 on tests B,C; Study 3 on C,D etc (Menten and Lesaffre, 2015, Nyaga et al., 2018b, Nyaga et al., 2018a, Lian et al., 2019, Ma et al., 2018, Owen et al., 2018).

8.2.5 Studies without a 'gold standard'

Models proposed by Lian et al. (2019) and Ma et al. (2018) are somewhat more flexible than others in that they allow for the possibility that some studies in the network of evidence did not apply the 'gold standard' test. The model of Menten and Lesaffre (2015) is more flexible still: it does not require any of the studies to have applied the gold standard. These models draw on the latent class modelling literature (see 7.3.3).

8.3 Recommendations

8.3.1 Guidance on synthesis of a single test

For synthesis of data on a singular diagnostic test, in the presence of a gold standard, the bivariate and HSROC models (Reitsma et al., 2005, Chu and Cole, 2006, Rutter and Gatsonis, 2001), recommended in the DAP manual, should continue to be recommended, although this advice could be made stronger now that these models are better established. Guidance could be made clearer, however, through provision of WinBUGS code and examples of use of synthesised results in a decision model. The guidance should further be extended, in particular to include models for multiple threshold data (Steinhauser et al., 2016, Jones et al., 2019) and identification of the optimal threshold in a decision model. As a first step, we recommend that a new TSD is prepared to cover these issues. This could be produced in the next 6 months.

8.3.2 Guidance on synthesis of comparative accuracy

In the longer term, we recommend that an additional TSD is prepared on methods for evidence synthesis of comparative and joint test accuracy. The methods described above are still in their infancy and require further testing before we are able to make clear recommendations. For this reason, a TSD on comparative and joint test accuracy will be feasible in 2 to 3 years, once further exploration and model development has taken place (see 7.3.4).

8.3.3 Tests with no gold standard

For completeness we note that there is also a need for guidance on evaluation of test accuracy in the absence of a 'gold standard' comparator*.* An approach often recommended is to evaluate tests through a Composite Reference Standard (Naaktgeboren et al., 2013), but this is based on assumptions that are recognised to be incoherent, and invariably produces biased estimates of sensitivity and specificity that depend on the true prevalence (Schiller et al., 2016). An alternative approach is use of latent class models. Latent class evidence synthesis models (de Bock et al., 1994, Walter et al., 1999, Dendukuri et al., 2012, Chu et al., 2009, Menten and Lesaffre, 2015, Menten et al., 2013, Liu et al., 2015) are currently underdeveloped, but this is an active area of research in several centres and appears to be a natural extension to comparative test accuracy models (see 7.2.5).

8.3.4 Research recommendations

Further research is required to test the robustness of the methods described above, using simulation studies and practical examples, in addition to further methodological developments, for comparative and joint test accuracy, and tests with no gold standard.

REFERENCES

ACHANA, F., COOPER, N., BUJKIEWICZ, S., HUBBARD, S., KENDRICK, D., JONES, D. & SUTTON, A. 2014. Network meta-analysis of multiple outcome measures accounting for borrowing of information across outcomes. *BMC Medical Research Methodology,* 14**,** 92 DOI: https://doi.org/10.1186/1471-2288-14-92.

ADES, A. E., LU, G., DIAS, S., MAYO-WILSON, E. & KOUNALI, D. 2015. Simultaneous synthesis of treatment effects and mapping to a common scale: an alternative to standardisation. *Research Synthesis Methods,* 6**,** 96-107 DOI: https://doi.org/10.1002/jrsm.1130

ALONSO A, T. B., T. BURZYKOWSKI, M. BUYSE, G. MOLENBERGHS, L. MUCHENE, N.J. PERUALILA, Z. SHKEDY, AND W. VAN DER ELST. 2016. *Applied Surrogate Endpoint Evaluation Methods with SAS and R.*, CRC Press DOI: https://www.crcpress.com/Applied-Surrogate-Endpoint-Evaluation-Methods-with-SAS-and-R/Alonso-Bigirumurame-Burzykowski-Buyse-Molenberghs-Muchene-Perualila-Shkedy-Elst/p/book/9781482249361

ANGLEMYER, A., HORVATH, H. T. & BERO, L. 2014. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database of Systematic Reviews,* DOI: https://doi.org/10.1002/14651858.MR000034.pub2.

ARENDS, L. R., HUNINK, M. G. M. & STIJNEN, T. 2008. Meta-analysis of summary survival curve data. *Statistics in Medicine,* 27**,** 4381-4396 DOI: https://doi.org/10.1002/sim.3311.

BALSHEM, H., HELFAND, M., SCHÜNEMANN, H. J., OXMAN, A. D., KUNZ, R., BROZEK, J., VIST, G. E., FALCK-YTTER, Y., MEERPOHL, J., NORRIS, S. & GUYATT, G. H. 2011. GRADE guidelines: 3. Rating the quality of evidence. *Journal of Clinical Epidemiology,* 64**,** 401-406 DOI: https://doi.org/10.1016/j.jclinepi.2010.07.015

BANBETA, A., VAN ROSMALEN, J., DEJARDIN, D. & LESAFFRE, E. 2019. Modified power prior with multiple historical trials for binary endpoints. *Statistics in Medicine,* 38**,** 1147-1169 DOI: https://doi.org/10.1002/sim.8019.

BARDEN, J., DERRY, S., MCQUAY, H. J. & MOORE, R. A. 2006. Bias from industry funding? A framework, a suggested approach, and a negative result. *Pain,* 121**,** 207-218 DOI: https://doi.org/10.1016/j.pain.2005.12.011.

BARRETT, J., FAREWELL, V., SIANNIS, F., TIERNEY, J. & HIGGINS, J. 2012. Two-stage meta-analysis of survival data from individual participants using percentile ratios. *Statistics in Medicine,* 31**,** 4296–4308 DOI: https://doi.org/10.1002/sim.5516.

BARTLETT, V. L., DHRUVA, S. S., SHAH, N. D., RYAN, P. & ROSS, J. S. 2019. Feasibility of Using Real-World Data to Replicate Clinical Trial Evidence. *JAMA Netw Open,* 2**,** e1912869 DOI: https://doi.org/10.1001/jamanetworkopen.2019.12869

BERKMAN, N. D., SANTAGUIDA, P. L., VISWANATHAN, M. & MORTON, S. C. 2014. *The empirical evidence of bias in trials measuring treatment differences,* Rockville (MD), Agency for Healthcare Research and Quality DOI: https://www.ncbi.nlm.nih.gov/pubmed/25392898.

BEXELIUS, C., QUIGLEY, J., THURESSON, P. & HAWKINS, N. 2014. The comparative efficacy of first-line (1L) treatments for stage IIIC and stage IV melanoma: results of a systematic review and networl meta-analysis. *Annals of Oncology,* 25**,** iv374–iv393 DOI: https://doi.org/10.1093/annonc/mdu344.11.

BHATTACHARJEE, A. 2019. A joint longitudinal and survival model for dynamic treatment regimes in Presence of Competing Risk Analysis. *Clinical Epidemiology and Global Health,* 7**,** 337-341 DOI: https://doi.org/10.1016/j.cegh.2018.09.001.

BOUCHER, R., ABRAMS, K. & LAMBERT, P. 2014. Simulating Individual Patient Level Data To Address Treatment Switching When Only Summary Data Are Available. *Value in Health,* 17**,** A579 DOI: https://doi.org/10.1016/j.jval.2014.08.1955.

BRAZIER, J. E., YANG, Y., TSUCHIYA, A. & ROWEN, D. L. 2010. A review of studies mapping (or cross-walking) non-preference based measures of health to generic preference-based measures. *European Journal of Health Economics,* 11**,** 215-225 DOI: https://doi.org/10.1007/s10198-009-0168-z.

BRIGNARDELLO-PETERSEN, R., BONNER, A., ALEXANDER, P., SIEMIENIUK, R., FURUKAWA, T., ROCHWERG, B., HAZLEWOOD, G., ALHAZZANI, W., MUSTAFA, R., MURAD, M., SCHÜNEMANN, H. & GUYATT, G. 2018. Advances in the GRADE approach to rate the certainty in estimates from a network meta-analysis. *Journal of Clinical Epidemiology,* 93**,** 36-44 DOI: https://doi.org/10.1016/j.jclinepi.2017.10.005.

BRIGNARDELLO-PETERSENA R, MUSTAFA RA, REED AC, SIEMIENIUKA M, MH, M., AGORITSAS T, IZCOVICH A, SCHÜNEMANN HJ, GUYATT GH & GROUP, G. W. 2019. GRADE approach to rate the certainty from a network meta-analysis: addressing incoherenc. *Journal of Clinical Epidemiology,* 109**,** 77-85 DOI: 10.1016/j.jclinepi.2018.11.025 https://www.sciencedirect.com/science/article/pii/S0895435618304967.

BUJKIEWICZ, S., ACHANA, F., PAPANIKOS, T., RILEY, R. D. & ABRAMS, K. R. 2019a. NICE DSU Technical Support Document 20: Multivariate meta-analysis of summary data for combining treatment effects on correlated outcomes and evaluating surrogate endpoints. ScHARR, University of Sheffield. DECISION SUPPORT UNIT, S., UNIVERSITY OF SHEFFIELD DOI: http://nicedsu.org.uk/wp-content/uploads/2019/10/TSD-20-mvmeta-final.pdf.

BUJKIEWICZ, S., JACKSON, D., THOMPSON, J. R., TURNER, R. M., STÄDLER, N., ABRAMS, K. R. & WHITE, I. R. 2019b. Bivariate network meta-analysis for surrogate endpoint evaluation. *Statistics in Medicine,* 38**,** 3322-3341 DOI: https://doi.org/10.1002/sim.8187.

BUJKIEWICZ, S., THOMPSON, J. R., RILEY, R. D. & ABRAMS, K. R. 2016. Bayesian meta-analytical methods to incorporate multiple surrogate endpoints in drug development process. *Statistics in Medicine,* 35**,** 1063-1089 DOI: https://doi.org/10.1002/sim.6776.

BUJKIEWICZ, S., THOMPSON, J. R., SPATA, E. & ABRAMS, K. R. 2017. Uncertainty in the Bayesian meta-analysis of normally distributed surrogate endpoints. *Statistical Methods in Medical Research,* 26**,** 2287-2318 DOI: https://doi.org/10.1177/0962280215597260.

BUJKIEWICZ, S., THOMPSON, J. R., SUTTON, A. J., COOPER, N. J., HARRISON, M. J., SYMMONS, D. P. & ABRAMS, K. R. 2014. Use of Bayesian multivariate meta-analysis to estimate the HAQ for mapping onto the EQ-5D questionnaire in rheumatoid arthritis. *Value in Health,* 17**,** 109-115 DOI: https://doi.org/10.1016/j.jval.2013.11.005.

BUJKIEWICZ, S., THOMPSON, J. R., SUTTON, A. J., COOPER, N. J., HARRISON, M. J., SYMMONS, D. P. M. & ABRAMS, K. R. 2013. Multivariate meta-analysis of mixed outcomes: a Bayesian approach. *Statistics in Medicine,* 32**,** 3926-3943 DOI: https://doi.org/10.1002/sim.5831

BURCH, J., PAULDEN, M., CONTI, S., STOCK, C., CORBETT, M., WELTON, N. J., ADES, A. E., SUTTON, A., COOPER, N., ELLIOT, A. J., NICHOLSON, K., DUFFY, S., MCKENNA, C., STEWART, S., WESTWOOD & PALMER, S. 2010. Antiviral drugs for the treatment of influenza: a systematic review and economic evaluation. *Health Technology Assesment,* 13**,** 1-290 DOI: https://www.ncbi.nlm.nih.gov/books/NBK78423/.

BURCH, J., PAULDEN, M., CONTI, S., STOCK, C., CORBETT, M., WELTON, N. J., ADES, A. E., SUTTON, A. J., COOPER, N., ELLIOT, N., NICHOLSON, K., DUFFY, S., MCKENNA, C., STEWART, L., WESTWOOD, M. & PALMER, S. 2008. Influenza - zanamivir, amantadine and oseltamivir (review): assessment report. NICE DOI: https://doi.org/10.3310/hta13580.

BURKE, D. L., BUJKIEWICZ, S. & RILEY, R. D. 2018. Bayesian bivariate meta-analysis of correlated effects: Impact of the prior distributions on the between-study correlation, borrowing of strength, and joint inferences. *Statistical Methods in Medical Research,* 27**,** 428-450 DOI: https://doi.org/10.1177/0962280216631361.

BURZYKOWSKI, T., MOLENBERGHS, G., BUYSE, M., GEYS, H. & RENARD, D. 2001. Validation of surrogate endpoints in mukltiple randomised clinical trials with failure-time endpoints. *Applied Statistics,* 50**,** 405-422 DOI: https://doi.org/10.1111/1467-9876.00244.

BURZYKOWSKI, T., MOLENBERGHS, G., AND BUYSE, M. 2006. *The evaluation of surrogate endpoints.*

, Springer.

BUYSE, M., MOLENBERGHS, G., BURZYKOWSKI, T., RENARD, D. & GEYS, H. 2000. The validation of surrogate endpoints in meta-analyses of randomised experiments. *Biostatistics,* 1**,** 49-67 DOI: https://doi.org/10.1093/biostatistics/1.1.49.

CALDWELL, D. M., ADES, A. E., DIAS, S., WATKINS, S., LI, T., TASKE, N., NAIDOO, B. & WELTON, N. J. 2016. A threshold analysis assessed the credibility of conclusions from network meta-analysis. *J Clin Epidemiol,* 80**,** 68-76 DOI: https://doi.org/10.1016/j.jclinepi.2016.07.003.

CARO, J. J. & ISHAK, K. J. 2010. No Head-to-Head Trial? Simulate the Missing Arms. *PharmacoEconomics,* 28**,** 957-967 DOI: https://doi.org/10.2165/11537420-000000000-00000.

CHAIMANI, A., VASILIADIS, H. S., PANDIS, N., SCHMID, C. H., WELTON, N. J. & SALANTI, G. 2013. Effects of study precision and risk of bias in networks of interventions: a network meta-epidemiological study. *International Journal Of Epidemiology,* 42**,** 1120-1131 DOI: https://doi.org/10.1093/ije/dyt074.

CHANG, S. M., MATCHAR, D. B., SMETANA, G. W., UMSCHEID, C. A., GRAY, R., TORCHIA, M. & ROCKVILLE, J. 2012. Methods guide for medical test reviews. *Agency for Healthcare Research Quality,* DOI: https://www.ncbi.nlm.nih.gov/books/NBK98241/.

CHENG, W. 2016. Network meta-analysis of diagnostic accuracy studies. *Unpublished PhD thesis. Brown University,* DOI: https://repository.library.brown.edu/studio/item/bdr:674079/.

CHU, H. T., CHEN, S. N. & LOUIS, T. A. 2009. Random Effects Models in a Meta-Analysis of the Accuracy of Two Diagnostic Tests Without a Gold Standard. *Journal of the American Statistical Association,* 104**,** 512-523 DOI: https://doi.org/10.1198/jasa.2009.0017

CHU, H. T. & COLE, S. R. 2006. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology,* 59**,** 1331-1332 DOI: https://doi.org/10.1016/j.jclinepi.2006.06.011

CIANI, O., BUYSE, M., DRUMMOND, M., RASI, G., SAAD, E. D. & TAYLOR, R. S. 2017. Time to Review the Role of Surrogate End Points in Health Policy: State of the Art and the Way Forward. *Value in Health,* 20**,** 487-495 DOI: https://doi.org/10.1016/j.jval.2016.10.011.

COLLETT, D. 2003. *Modelling survival data in medical research,* Boca raton. Florida, Chapman & Hall/CRC.

COPE, S., CHAN, K. & JANSEN, J. 2017. Multivariate Network Meta-Analysis of Survival Function Parameters. *Value in Health,* 20**,** A401-A402 DOI: https://doi.org/10.1016/j.jval.2017.08.020.

CORBETT, M., SOARES, M., JHUTI, G., RICE, S., SPACKMAN, E., SIDERIS, E., MOE-BYRNE, T., FOX, D., MARZO-ORTEGA, H., KAY, L., WOOLACOTT, N. & PALMER, S. 2016. Tumour necrosis factor-α inhibitors for ankylosing spondylitis and non-radiographic axial spondyloarthritis: a systematic review and economic evaluation. *Health Technology Assesment,* 20**,** 1-334 DOI: https://doi.org/10.3310/hta20090.

DAKIN, H., FIDLER, C. & HARPER, C. 2010. Mixed Treatment Comparison Meta-Analysis Evaluating the Relative Efficacy of Nucleos(t)ides for Treatment of Nucleos(t)ide-Naive Patients with Chronic Hepatitis B. *Value in Health,* 13**,** 934-945 DOI: htpps://doi.org/10.1111/j.1524-4733.2010.00777.x.

DAKIN, H. A., WELTON, N. J., ADES, A. E., COLLINS, S., ORME, M. & KELLY, S. 2011. Mixed treatment comparison of repeated measurements of a continuous endpoint: an example using topical treatments for primary open-angle glaucoma and ocular hypertension. *Statistics In Medicine,* 30**,** 2511-2535 DOI: https://doi.org/10.1002/sim.4284.

DANIELS, M. J. & HUGHES, M. D. 1997. Meta-analysis for the evaluation of potential surrogate markers. *Statistics In Medicine,* 16**,** 1965-1982 DOI: https://doi.org/10.1002/(SICI)1097-0258(19970915)16:17<1965::AID-SIM630>3.0.CO;2-M.

DE BOCK, G. H., HOUWING-DUISTERMAAT, J. J., SPRINGER, M. P., KIEVIT, J. & VAN HOUWELINGEN, J. C. 1994. Sensitivity and specificity of diagnostic tests in acute maxillary sinusitis determined by maximum likelihood in the absence of an external standard. *J Clin Epidemiol,* 47**,** 1343-52 DOI: https://doi.org/10.1016/0895-4356(94)90078-7

DEAR, K. G. B. 1994. Iterative generalised least squares for meta-analysis of survival data at multiple times. *Biometrics,* 50**,** 989-1982.

DEL GIOVANI, C., VACCHI, L., MAVRIDIS, D., FILIPPINI, G. & SALANTI, G. 2013. Network meta-analysis models to account for variability in treatment definitions: application to dose effects. *Statistics in Medicine,* 32**,** 25-39 DOI: https://doi.org/10.1002/sim.5512.

DEMIRIS, N. & SHARPLES, L. D. 2006. Bayesian evidence synthesis to extrapolate survival estimates in cost-effectiveness studies. *Statistics in Medicine,* 25**,** 1960-1975 DOI: https://doi.org/10.1002/sim.2366.

DENDUKURI, N., SCHILLER, I., JOSEPH, L. & PAI, M. 2012. Bayesian Meta-Analysis of the Accuracy of a Test for Tuberculous Pleuritis in the Absence of a Gold Standard Reference. *Biometrics,* 68**,** 1285-1293 DOI: https://doi.org/10.1111/j.1541-0420.2012.01773.x

DENIZ, B., ALTINCATAL, A., AMBAVANE, A., RAO, S., DOAN, J., MALCOLM, B. & AL, E. 2018. Application of dynamic modeling for survival estimation in advanced renal cell carcinoma. *PLoS ONE* 13**,** e0203406 DOI: https://doi.org/10.1371/journal.pone.0203406.

DIAS, S. & ADES, A. E. 2016. Absolute or relative effects? Arm-based synthesis of trial data. *Res Synth Methods,* 7**,** 23-8 DOI: https://doi.org/10.1002/jrsm.1184.

DIAS, S., ADES, A. E., WELTON, N. J., JANSEN, J. P. & SUTTON, A. J. 2018. *Network meta-analysis for decision making*, Wiley DOI: https://www.wiley.com/en-gb/Network+Meta+Analysis+for+Decision+Making-p-9781118647509.

DIAS, S., SUTTON, A. J., ADES, A. E. & WELTON, N. J. 2013. Evidence Synthesis for Decision Making 2: A generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making,* 33**,** 607-617 DOI: https://doi.org/10.1177/0272989X12458724

DIAS, S., SUTTON, A. J., WELTON, N. J. & ADES, A. E. 2011a. NICE DSU Technical Support Document 3: Heterogeneity: subgroups, meta-regression, bias and bias-adjustment. Technical Support Document. UNIT, N. D. S. DOI: http://nicedsu.org.uk/wp-content/uploads/2016/03/TSD3-Heterogeneity.final-report.08.05.12.pdf.

DIAS, S., WELTON, N. J. & ADES, A. E. 2010a. Study designs to detect sponsorship and other biases in systematic reviews. *Journal of Clinical Epidemiology,* 63**,** 587-588 DOI: https://doi.org/10.1016/j.jclinepi.2010.01.005.

DIAS, S., WELTON, N. J., CALDWELL, D. M. & ADES, A. E. 2010b. Checking Consistency in Mixed Treatment Comparison Meta-analysis. *Statistics In Medicine,* 29**,** 932-944 DOI: https://doi.org/10.1002/sim.3767.

DIAS, S., WELTON, N. J., MARINHO, V. C. C., SALANTI, G., HIGGINS, J. P. T. & ADES, A. E. 2010c. Estimation and adjustment of bias in randomised evidence by using Mixed Treatment Comparison Meta-analysis. *Journal of the Royal Statistical Society (A),* 173**,** 613-629 DOI: https://doi.org/10.1111/j.1467-985X.2010.00639.x.

DIAS, S., WELTON, N. J., SUTTON, A. J. & ADES, A. E. 2011b. NICE DSU Technical Support Document 2: A generalised linear modelling framework for pair-wise and network meta-analysis of randomised controlled trials. Technical Support Document. UNIT, N.

D. S. DOI: http://nicedsu.org.uk/technical-support-documents/evidence-synthesis-tsd-series/.

DIAS, S., WELTON, N. J., SUTTON, A. J., CALDWELL, D. M., LU, G. & ADES, A. E. 2011c. NICE DSU Technical Support Document 4: Inconsistency in networks of evidence based on randomised controlled trials. Technical Support Document. UNIT, N. D. S. DOI: http://nicedsu.org.uk/technical-support-documents/evidence-synthesis-tsd-series/.

DIMOU, N. L., ADAM, M. & BAGOS, P. G. 2016. A multivariate method for meta-analysis and comparison of diagnostic tests. *Stat Med,* 35**,** 3509-23 DOI: https://doi.org/10.1002/sim.6919

DOMINICI, F., PARMIGIANI, G., WOLPERT, R. L. & HASSELBLAD, V. 1999. Meta-analysis of migraine headache treatments: combining information from heterogenous designs. *Journal Of The American Statistical Association,* 94**,** 16-28.

DUAN, Y., SMITH, E. P. & YE, K. 2006. Using power priors to improve the binomial test of water quality. *Journal of Agricultural, Biological, and Environmental Statistics,* 11**,** 151 DOI: https://doi.org/10.1198/108571106X110919.

EFTHIMIOU, O., MAVRIDIS, D., CIPRIANI, A., LEUCHT, S., BAGOS, P. & SALANTI, G. 2014. An approach for modelling multiple correlated outcomes in a network of interventions using odds ratios. *Statistics in Medicine,* 33**,** 2275-2287 DOI: https://doi.org/10.1002/sim.6117.

EFTHIMIOU, O., MAVRIDIS, D., DEBRAY, T. P. A., SAMARA, M., BELGER, M., SIONTIS, G. C. M., LEUCHT, S., SALANTI, G. & ON BEHALF OF GETREAL WORK PACKAGE 4 2017. Combining randomized and non-randomized evidence in network meta-analysis. *Statistics in Medicine,* 36**,** 1210-1226 DOI: https://doi.org/10.1002/sim.7223.

EFTHIMIOU, O., MAVRIDIS, D., RILEY, R., CIPRIANI, A. & SALANTI, G. 2015. Joint synthesis of multiple correlated outcomes in networks of interventions. *Biostatistics,* 16**,** 84-97 DOI: https://doi.org/10.1093/biostatistics/kxu030.

ELIA, E. G., STÄDLER, N., CIANI, O., TAYLOR, R. S. & BUJKIEWICZ, S. 2020. Combining tumour response and progression free survival as surrogate endpoints for overall survival in advanced colorectal cancer. *Cancer Epidemiology,* 64**,** 101665 DOI: https://doi.org/10.1016/j.canep.2019.101665.

FIOCCO, M., PUTTER, H. & VAN HOUWELINGEN, J. C. 2009. Meta-analysis of pairs of survival curves under heterogeneity: A Poisson correlated gamma-frailty approach. *Statistics in Medicine,* 28**,** 3782-3797 DOI: https://doi.org/10.1002/sim.3752.

FLACCO, M. E., MANZOLI, L., BOCCIA, S., CAPASSO, L., ALEKSOVSKA, K., ROSSO, A., SCAIOLI, G., DE VITO, C., SILIQUINI, R., VILLARI, P. & IOANNIDIS, J. P. A. 2015. Head-to-head randomized trials are mostly industry sponsored and almost always favor the industry sponsor. *Journal of Clinical Epidemiology,* 68**,** 811-820 DOI: https://doi.org/10.1016/j.jclinepi.2014.12.016.

FLEMING, T. R. & DEMETS, D. L. 1996. Surrogate end-points in clinical trials: are we being misled? *Annals of Internal Medicine,* 125**,** 605-613 DOI: https://doi.org/10.7326/0003-4819-125-7-199610010-00011.

FRANCHINI, A., DIAS, S., ADES, A. E., JANSEN, J. & WELTON, N. 2012. Accounting for correlation in mixed treatment comparisons with multi-arm trials. *Research Synthesis Methods,* 3**,** 142-160 DOI: https://doi.org/10.1002/jrsm.1049.

FREEMAN, S. & CARPENTER, J. 2017. Bayesian one-step IPD network meta-analysis of time-to-event data using Royston-Parmar models. *Research Synthesis Methods,* 8**,** 451-464 DOI: https://doi.org/10.1002/jrsm.1253.

FRIEDMAN, M. 1982. Piecewise exponential models for survival data with covariates. *Annals of Statistics,* 10**,** 101-113 DOI: https://projecteuclid.org/euclid.aos/1176345693.

FUNK, M., WESTREICH, D., WIESEN, C., STÜRMER, T., BROOKHART, M. & DAVIDIAN, M. 2011. Doubly Robust Estimation of Causal Effects. *American Journal of Epidemiology,* 173**,** 761-767 DOI: https://doi.org/10.1093/aje/kwq439.

GARTLEHNER, G. & FLEG, A. 2010. Pharmaceutical company-sponsored drug trials: the system is broken. *Journal of Clinical Epidemiology,* 63**,** 128-129 DOI: https://doi.org/10.1016/j.jclinepi.2009.07.015.

GARTLEHNER, G., MORGAN, L., THIEDA, P. & FLEG, A. 2010. The effect of sponsorship on a systematically evaluated body of evidence of head-to-head trials was modest: secondary analysis of a systematic review. *Journal of Clinical Epidemiology,* 63**,** 117-125 (doi:10.1016/j.jclinepi.2008.09.019) DOI: https://doi.org/10.1016/j.jclinepi.2008.09.019.

GELMAN, A. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis,* 1**,** 515-533 DOI: https://projecteuclid.org/euclid.ba/1340371048.

GUYOT, P., ADES, A. E., BEASLEY, M., LUEZA, B., PIGNON, J.-P. & WELTON, N. J. 2016. Extrapolation of survival curves from cancer trials using external information. *Medical Decision Making,* 37**,** 353-366 DOI: https://doi.org/10.1177/0272989x16670604.

GUYOT, P., ADES, A. E., OUWENS, M. J. N. M. & WELTON, N. J. 2012. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology,* 12**,** 9 DOI: https://doi.org/10.1186/1471-2288-12-9.

GUYOT, P., WELTON, N. J., OUWENS, J. N. M. & ADES, A. E. 2011. Survival time outcomes in randomised controlled trials and meta-analyses: the parallel universes of efficacy and cost-effectiveness. . *Value in Health,* 14**,** 640-646 DOI: https://doi.org/10.1177/0272989X16670604.

HAAS, D. M., CALDWELL, D. M., KIRKPATRICK, P., MCINTOSH, J. J. & WELTON, N. J. 2012. Tocolytic therapy for preterm delivery: systematic review and network meta-analysis. *British Medical Journal,* 345 DOI: https://doi.org/10.1136/bmj.e6226.

HENSHALL, C., LATIMER, N., SANSOM, L. & WARD, R. 2016. Treatment switching in cancer trials: issues and proposals. *International Journal of Techology Assessment and Health Care,* 32**,** 167-174 DOI: https://doi.org/10.1017/S026646231600009X.

HERNÁN, M. & ROBINS, J. 2020. *Causal Inference: What If. ,* Boca Raton, Chapman & Hall/CRC DOI: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

HIGGINS, J. P., JACKSON, D., BARRETT, J. K., LU, G., ADES, A. E. & WHITE, I. R. 2012. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Res Synth Methods,* 3**,** 98-110 DOI: https://doi.org/10.1002/jrsm.1044.

HIGGINS, J. P. T., WHITE, I. R. & WOOD, A. 2008. Imputation methods for missing outcome data in meta-analysis of clinical trials. *Clinical Trials,* 5**,** 225-239 DOI: https://doi.org/10.1177/1740774508091600.

HIGGINS, J. P. T. & WHITEHEAD, A. 1996. Borrowing strength from external trials in a meta-analysis. *Statistics In Medicine,* 15**,** 2733-2749 DOI: https://doi.org/10.1002/(SICI)1097-0258(19961230)15:24<2733::AID-SIM562>3.0.CO;2-0.

HOBBS, B. P., CARLIN, B. P., MANDREKAR, S. J. & SARGENT, D. J. 2011. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics,* 67**,** 1047-56 DOI: https://doi.org/10.1111/j.1541-0420.2011.01564.x.

HOBBS, B. P., CARLIN, B. P. & SARGENT, D. J. 2013. Adaptive adjustment of the randomization ratio using historical control data. *Clin Trials,* 10**,** 430-40 DOI: https://doi.org/10.1177/1740774513483934.

HOBBS, B. P., SARGENT, D. J. & CARLIN, B. P. 2012. Commensurate Priors for Incorporating Historical Information in Clinical Trials Using General and Generalized Linear Models. *Bayesian Anal,* 7**,** 639-674 DOI: https://doi.org/10.1214/12-ba722.

HONG, H., CHU, H., ZHANG, J. & CARLIN, B. P. 2016a. A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Research Synthesis Methods,* 7**,** 6-22 DOI: https://doi.org/10.1002/jrsm.1153.

HONG, H., CHU, H., ZHANG, J. & CARLIN, B. P. 2016b. Rejoinder to the discussion of "a Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons," by S. Dias and A. E. Ades. *Research Synthesis Methods,* 7**,** 29-33 DOI: https://doi.org/10.1002/jrsm.1186.

HONG, H., FU, H. & CARLIN, B. P. 2018. Power and commensurate priors for synthesizing aggregate and individual patient level data in network meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 67**,** 1047-1069 DOI: https://doi.org/10.1111/rssc.12275.

HONG, H., FU, H., PRICE, K. L. & CARLIN, B. P. 2015. Incorporation of individual-patient data in network meta-analysis for multiple continuous endpoints, with application to diabetes treatment. *Statistics in Medicine,* 34**,** 2794-2819 DOI: https://doi.org/10.1002/sim.6519.

HOYER, A. & KUSS, O. 2018a. Meta-analysis for the comparison of two diagnostic tests to a common gold standard: A generalized linear mixed model approach. *Statistical Methods in Medical Research,* 27**,** 1410-1421 DOI: https://doi.org/10.1177/0962280216661587

HOYER, A. & KUSS, O. 2018b. Meta-analysis for the comparison of two diagnostic tests-A new approach based on copulas. *Statistics in Medicine,* 37**,** 739-748 DOI: https://doi.org/10.1002/sim.7556

HOYLE, M. W. & HENLEY, W. 2011. Improved curve fits to summary survival data: application to economic evaluation of health technologies. *BMC Medical Research Methodology,* 11**,** 1-14 DOI: https://doi.org/10.1186/1471-2288-11-139.

HULTCRANTZ, M., RIND, D., AKL, E. A., TREWEEK, S., MUSTAFA, R. A., IORIO, A., ALPER, B. S., MEERPOHL, J. J., MURAD, M. H., ANSARI, M. T., KATIKIREDDI, S. V., OSTLUND, P., TRANAEUS, S., CHRISTENSEN, R., GARTLEHNER, G., BROZEK, J., IZCOVICH, A., SCHUNEMANN, H. & GUYATT, G. 2017. The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol,* 87**,** 4-13 DOI: https://doi.org/10.1016/j.jclinepi.2017.05.006.

HWANG, H. & DESANTIS, S. M. 2018. Multivariate network meta-analysis to mitigate the effects of outcome reporting bias. *Statistics in Medicine,* 37**,** 3254-3266 DOI: https://doi.org/10.1002/sim.7815.

IBRAHIM, J. G. & CHEN, M.-H. 2000. Power prior distributions for regression models. *Statist. Sci.,* 15**,** 46-60 DOI: https://doi.org/10.1214/ss/1009212673.

INSTITUTE OF SOCIAL AND PREVENTIVE MEDICINE UNIVERSITY OF BERNE. *CINeMA: Confidence in Network Meta-analysis [software]* [Online]. Available: www.cinema.ispm.ch [Accessed February 20, 2020].

IOANNIDIS, J. P. A., HAIDICH, A.-B., PAPPA, M., PANTZAKIS, N., KOKORI, S. I., TEKTONIDOU, M. G., CONTOPOULOS-IOANNIDIS, D. G. & LAU, J. 2001. Comparison of evidence of treatment effects in randomised and non-randomised studies. *JAMA,* 286**,** 821-830 DOI: https://doi.org/10.1001/jama.286.7.821.

ISHAK, J. 2014. Indirect Treatment Comparison Without Network Meta-Analysis: Overview of Novel Techniques. Evidera.

ISHAK, K., CARO, J., DRAYSON, M., DIMOPOULOS, M., WEBER, D., AUGUSTSON, B., CHILD, J., KNIGHT, R., IQBAL, G., DUNN, J., SHEARER, A. & MORGAN, G. 2011. Adjusting for patient crossover in clinical trials using external data: a case study of lenalidomide for advanced multiple myeloma. *Value in Health,* 14**,** 672-678 DOI: https://doi.org/10.1016/j.jval.2011.02.1182.

ISHAK, K. J., PLATT, R. W., JOSEPH, L. & HANLEY, J. A. 2008. Impact of approximating or ignoring within-study covariances in multivariate meta-analyses. *Statistics in Medicine,* 27**,** 670-686 DOI: https://doi.org/10.1002/sim.2913.

JACKSON, C., STEVENS, J., REN, S., LATIMER, N., BOJKE, L., MANCA, A. & SHARPLES, L. 2017. Extrapolating Survival from Randomized Trials Using External Data: A Review of Methods. *Medical Decision Making,* 37**,** 377-390 DOI: https://doi.org/10.1177/0272989X16639900.

JACKSON, D., BARRETT, J. K., RICE, S., WHITE, I. R. & HIGGINS , J. P. T. 2014. A design-by-treatment interaction model for network meta-analysis with random inconsistency effects. *Statistics in Medicine,* 33**,** 3639-3654 DOI: https://doi.org/10.1002/sim.6188.

JACKSON, D., BUJKIEWICZ, S., LAW, M., RILEY, R. D. & WHITE, I. R. 2018. A matrix-based method of moments for fitting multivariate network meta-analysis models with multiple outcomes and random inconsistency effects. *Biometrics,* 74**,** 548-556 DOI: https://doi.org/10.1111/biom.12762.

JACKSON, D., RILEY, R. & WHITE, I. R. 2011. Multivariate meta-analysis: potential and promise. *Statistics In Medicine,* 30**,** 2481-2598 DOI: https://doi.org/10.1002/sim.4172.

JANSEN, J. & TRIKALINOS, T. 2013. Multivariate Network Meta-Analysis of Progression Free Survival and Overall Survival. *Valuein Health,* 16**,** A617 DOI: https://doi.org/10.1016/j.jval.2013.08.1791.

JANSEN, J. P. 2011. Network meta-analysis of survival data with fractional polynomials. *BMC Medical Research Methodology,* 11**,** 1-14 DOI: https://doi.org/10.1186/1471-2288-11-61.

JANSEN, J. P., VIEIRA, M. C. & COPE, S. 2015. Network meta-analysis of longitudinal data using fractional polynomials. *Statistics in Medicine,* 34**,** 2294-2311 DOI: https://doi.org/10.1002/sim.6492

JONES, H. E., GATSONSIS, C. A., TRIKALINOS, T. A., WELTON, N. J. & ADES, A. E. 2019. Quantifying how diagnostic test accuracy depends on threshold in a meta-analysis. *Stat Med,* DOI: https://doi.org/10.1002/sim.8301

KEENEY, E., DAWOUD, D. & DIAS, S. 2018. Different Methods for Modelling Severe Hypoglycaemic Events: Implications for Effectiveness, Costs and Health Utilities. *PharmacoEconomics,* DOI: https://doi.org/10.1007/s40273-018-0612-y.

KEW, K. M., DIAS, S. & CATES, C. J. 2014. Long-acting inhaled therapy (beta-agonists, anticholinergics and steroids) for COPD: a network meta-analysis. *Cochrane Database of Systematic Reviews***,** Art No.: CD010844 DOI: https://doi.org/10.1002/14651858.CD010844.pub2.

KIRKHAM, J. J., RILEY, R. D. & WILLIAMSON, P. R. 2012. A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews. *Statistics in Medicine,* 31**,** 2179-2195 DOI: https://doi.org/10.1002/sim.5356.

KOUNALI, D. Z., BUTTON, K. S., LEWIS, G. & ADES, A. E. 2016. The relative responsiveness of test instruments can be estimated using a meta-analytic approach: an illustration with treatments for depression. *Journal of Clinical Epidemiology,* 77**,** 68-77 DOI: https://doi.org/10.1016/j.jclinepi.2016.03.005.

KRAHN, U., BINDER, H. & KONIG, J. 2013. A graphical tool for locating inconsistency in network meta-analyses. *BMC Medical Research Methodology,* 13**,** 35 DOI: https://doi.org/10.1186/1471-2288-13-35.

KRISHAN, A., WELTON, N., DWAN, K., SUDELL, M. & TUDUR-SMITH, C. Review of reporting of time to event analyses and the proportional hazards assumption in meta-analysis. International Society for Clinical Biostatistics, 14th-18th July 2019 2019 Leuven, Belgium.

LASSERE, M. N., JOHNSON, K. R., SCHIFF, M. & REES, D. 2012. Is blood pressure reduction a valid surrogate endpoint for stroke prevention? an analysis incorporating a systematic review of randomised controlled trials, a by-trial weighted errors-in-variables regression, the surrogate threshold effect (STE) and the biomarker-surrogacy (BioSurrogate) evaluation schema (BSES). *BMC Medical Research Methodology,* 12**,** 27 DOI: https://doi.org/10.1186/1471-2288-12-27.

LATIMER, N. 2011. NICE DSU Technical Support Document 14: Undertaking survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data. NICE Decision Support Unit. DOI: http://nicedsu.org.uk/technical-support-documents/survival-analysis-tsd/.

LATIMER, N., WHITE, I., ABRAMS, K. & SIEBERT, U. 2019. Causal inference for long-term survival in randomised trials with treatment switching: Should re-censoring be applied when estimating counterfactual survival times? *Statistical Methods in Medical Research,* 28**,** 2475-2493 DOI: https://doi.org/10.1177/0962280218780856.

LATIMER, N., WHITE, I., TILLING, K. & SIEBERT, U. 2020. Improved two-stage estimation to adjust for treatment switching in randomised trials: g-estimation to address time-dependent confounding. *Statistical Methods in Medical Research,* DOI: https://doi.org/10.1177/0962280220912524

LATIMER, N. R. & ABRAMS, K. R. 2014. NICE DSU Technical Support Document 16: Adjusting survival time estimates n the presence of treatment switching. NICE Decision Support Unit. UNIT, N. D. S. DOI: http://nicedsu.org.uk/wp-content/uploads/2016/03/TSD16_Treatment_Switching.pdf.

LIAN, Q. S., HODGES, J. S. & CHU, H. T. 2019. A Bayesian Hierarchical Summary Receiver Operating Characteristic Model for Network Meta-Analysis of Diagnostic Tests. *Journal of the American Statistical Association,* 114**,** 949-961 DOI: https://doi.org/10.1080/01621459.2018.1476239

LIU, Y. L., CHEN, Y. & CHU, H. T. 2015. A unification of models for meta-analysis of diagnostic accuracy studies without a gold standard. *Biometrics,* 71**,** 538-547 DOI: https://doi.org/10.1111/biom.12264

LU, G. & ADES, A. 2006. Assessing evidence consistency in mixed treatment comparisons. *Journal Of The American Statistical Association,* 101**,** 447-459 DOI: https://doi.org/10.1198/016214505000001302.

LU, G. & ADES, A. 2009. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics,* 10**,** 792-805 DOI: https://doi.org/10.1093/biostatistics/kxp032.

LU, G., ADES, A. E., SUTTON, A. J., COOPER, N. J., BRIGGS, A. H. & CALDWELL, D. M. 2007. Meta-analysis of mixed treatment comparisons at multiple follow-up times. *Statistics In Medicine,* 26**,** 3681-3699 DOI: https://doi.org/10.1002/sim.2831.

LU, G., KOUNALI, D. & ADES, A. E. 2014. Simultaneous multi-outcome synthesis and mapping of treatment effects to a common scale. *Value in Health,* 17**,** 280-287 DOI: https://doi.org/10.1016/j.jval.2013.12.006.

LUNN, D., JACKSON, C., BEST, N., THOMAS, A. & SPIEGELHALTER, D. 2013. *The BUGS book,* Boca Raton, FL, CRC Press DOI: https://www.crcpress.com/The-BUGS-Book-A-Practical-Introduction-to-Bayesian-Analysis/Lunn-Jackson-Best-Thomas-Spiegelhalter/p/book/9781584888499.

MA, X. Y., LIAN, Q. S., CHU, H. T., IBRAHIM, J. G. & CHEN, Y. 2018. A Bayesian hierarchical model for network meta-analysis of multiple diagnostic tests. *Biostatistics,* 19**,** 87-102 DOI: https://doi.org/10.1093/biostatistics/kxx025

MACASKILL, P., GATSONIS, C., DEEKS, J., HARBORD, R. & TAKWOINGI, Y. 2010. Chapter 10: Analysing and presenting results. *In:* DEEKS, J. J., BOSSUYT, P. M. & GATSONIS, C. (eds.) *Cochrane handbook for systematic reviews of diagnostic test accuracy.* London: The Cochrane Collaboration.

MAVRIDIS, D., CHAIMANI, A., EFTHIMIOU, O., LEUCHT, S. & SALANTI, G. 2014. Addressing missing outcome data in meta-analysis. *Evidence Based Mental Health,* 17**,** 85-89 DOI: https://doi.org/10.1136/eb-2014-101900.

MAVRIDIS, D., SALANTI, G., FURUKAWA, T. A., CIPRIANI, A., CHAIMANI, A. & WHITE, I. R. 2019. Allowing for uncertainty due to missing and LOCF imputed outcomes in meta-analysis. *Statistics in Medicine,* 38**,** 720-737 DOI: 10.1002/sim.8009 https://onlinelibrary.wiley.com/doi/full/10.1002/sim.8009.

MAWDSLEY, D., BENNETTS, M., DIAS, S., BOUCHER, M. & WELTON, N. J. 2016. Model-Based Network Meta-Analysis: A Framework for Evidence Synthesis of Clinical Trial Data. *CPT Pharmacometrics Syst Pharmacol,* 5**,** 393-401 DOI: https://doi.org/10.1002/psp4.12091.

MAYO-WILSON, E., DIAS, S., MAVRANEZOULI, I., KEW, K., CLARK, D. M., ADES, A. E. & PILLING, S. 2014. Psychological and pharmacological interventions for social anxiety disorder in adults: a systematic review and network meta-analysis. *Lancet Psychiatry,* 1**,** 368-376 DOI: http://dx.doi.org/10.1016/S2215-0366(14)70329-3.

MELENDEZ-TORRE, G. J., BONELL, C. & THOMAS, J. 2015. Emergent approaches to the meta-analysis of multiple heterogeneous complex interventions. *BMC Medical Research Methodology,* 15**,** 47 DOI: https://doi.org/10.1186/s12874-015-0040-z.

MENTEN, J., BOELAERT, M. & LESAFFRE, E. 2013. Bayesian meta-analysis of diagnostic tests allowing for imperfect reference standards. *Statistics in Medicine,* 32**,** 5398-5413 DOI: https://doi.org/10.1002/sim.5959

MENTEN, J. & LESAFFRE, E. 2015. A general framework for comparative Bayesian meta-analysis of diagnostic studies. *Bmc Medical Research Methodology,* 15 DOI: https://doi.org/10.1186/s12874-015-0061-7

MILLS, E. J., THORLUND, K. & IOANNIDIS, J. P. A. 2012. Calculating additive treatment effects from multiple randomized trials provides useful estimates of combination therapies. *Journal of Clinical Epidemiology,* 65**,** 1282-1288 DOI: https://doi.org/10.1016/j.jclinepi.2012.07.012.

MORENO, S. G., SUTTON, A. J., ADES, A. E., STANLEY, T. D., ABRAMS, K. R., PETERS, J. L. & COOPER, N. J. 2009a. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology,* 9 DOI: https://doi.org/10.1186/1471-2288-9-2.

MORENO, S. G., SUTTON, A. J., TURNER, E. H., ABRAMS, K. R., COOPER, N. J., PALMER, T. M. & ADES, A. E. 2009b. Novel methods to deal with publication biases: secondary analysis of antidepressant trials in the FDA trial registry database and related journal publications. *BMJ,* 339**,** b2981 DOI: https://doi.org/10.1136/bmj.b2981.

MOUSTGAARD, H., CLAYTON, G. L., JONES, H. E., BOUTRON, I., JØRGENSEN, L., LAURSEN, D. R. T., OLSEN, M. F., PALUDAN-MÜLLER, A., RAVAUD, P., SAVOVIĆ, J., STERNE, J. A. C., HIGGINS, J. P. T. & HRÓBJARTSSON, A. 2020. Impact of blinding on estimated treatment effects in randomised clinical trials: meta-epidemiological study. *BMJ,* 368**,** l6802 DOI: https://doi.org/10.1136/bmj.l6802.

NAAKTGEBOREN, C. A., BERTENS, L. C., VAN SMEDEN, M., DE GROOT, J. A., MOONS, K. G. & REITSMA, J. B. 2013. Value of composite reference standards in diagnostic research. *BMJ,* 347**,** f5605 DOI: https://doi.org/10.1136/bmj.f5605.

NACI, H., DIAS, S. & ADES, A. E. 2014. Industry sponsorship bias in research findings: A network meta-analytic exploration of LDL cholesterol reduction in the randomised trials of statins *BMJ,* 349**,** g5741 DOI: http://dx.doi.org/10.1136/bmj.g5741.

NATIONAL CLINICAL GUIDELINE CENTRE. 2012. Crohn's disease. Management in adults, children and young people. EXCELLENCE, N. I. F. H. A. C. DOI: https://www.ncbi.nlm.nih.gov/pubmed/25340220.

NATIONAL COLLABORATING CENTRE FOR MENTAL HEALTH. 2014. The assessment and management of bipolar disorder in adults, children and young people in primary and secondary care, Updated edition

. National Clinical Guideline Number 185. National Institute for Health and Care Excellence. PSYCHIATRISTS, T. B. P. S. A. T. R. C. O. DOI: https://www.ncbi.nlm.nih.gov/pubmed/29718639.

NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE. 2012. Roflumilast for the management of severe chronic obstructive pulmonary disease. NICE technology appraisal guidance.

NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE. 2013. Hyperphosphataemia in chronic kidney disease. NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE DOI: https://www.nice.org.uk/guidance/cg157.

NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE. 2017. Eating disorders: recognition and treatment. National Institute for Health and Care Excellence. EXCELLENCE, N. I. F. H. A. C. DOI: https://www.nice.org.uk/guidance/NG69.

NICE. 2011. *Diagnostic Assessment Programme manual* [Online]. Available: https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-diagnostics-guidance/Diagnostics-assessment-programme-manual.pdf [Accessed].

NIKOLAKOPOULOU, N., HIGGINS, J. P. T., PAPAKONSTANTINOU, T., CHAIMANI, A., DEL GIOVANE, C., EGGER, M. & SALANTI, G. 2019. Assessing Confidence in the Results of Network Meta-Analysis (Cinema). *bioRxiv***,** 57 DOI: 10.1101/597047 https://www.biorxiv.org/content/10.1101/597047v1.

NOVIELLI, N., SUTTON, A. & COOPER, N. 2013. Meta-analysis of the accuracy of two diagnostic tests used in combination: application to the ddimer test and the wells score for the diagnosis of deep vein thrombosis. *Value in Health,* 16**,** 619-628 DOI: https://doi.org/10.1016/j.jval.2013.02.007.

NYAGA, V. N., AERTS, M. & ARBYN, M. 2018a. ANOVA model for network meta-analysis of diagnostic test accuracy data. *Statistical Methods in Medical Research,* 27**,** 1766-1784 DOI: https://doi.org/10.1177/0962280216669182

NYAGA, V. N., ARBYN, M. & AERTS, M. 2018b. Beta-binomial analysis of variance model for network meta-analysis of diagnostic test accuracy data. *Statistical Methods in Medical Research,* 27**,** 2554-2566 DOI: https://doi.org/10.1177/0962280216682532

O'HAGAN, A., BUCK, C. E., ALIREZA DANESHKHAH, J., EISER, R., GARTHWAITE, P. H., JENKINSON, D. J., OAKLEY, J. E. & RAKOW, T. 2006. *Uncertain judgements: eliciting experts' probabilities*, Wiley DOI: https://www.wiley.com/en-gb/Uncertain+Judgements%3A+Eliciting+Experts%27+Probabilities-p-9780470029992.

OGANISIAN, A. & ROY, J. 2020. A Practical Introduction to Bayesian Estimation of Causal Effects: Parametric and Nonparametric Approaches. *ArXiv,* DOI: https://arxiv.org/pdf/2004.07375.pdf.

OHLSSEN, D., PRICE, K. L., XIA, H. A., HONG, H., KERMAN, J., FU, H., QUARTEY, G., HEILMANN, C. R., MA, H. & CARLIN, B. P. 2014. Guidance on the implementation and reporting of a drug safety Bayesian network meta-analysis. *Pharmaceutical Statistics,* 13**,** 55-70 DOI: https://doi.org/10.1002/pst.1592.

OUWENS, M. J. N. M., PHILIPS, Z. & JANSEN, J. P. 2010. Network meta-analysis of parametric survival curves. *Research Synthesis Methods,* 1**,** 258-271 DOI: https://doi.org/10.1002/jrsm.25.

OWEN, R. K., COOPER, N. J., QUINN, T. J., LEES, R. & SUTTON, A. J. 2018. Network meta-analysis of diagnostic test accuracy studies identifies and ranks the optimal diagnostic tests and thresholds for health care policy and decision-making. *Journal of Clinical Epidemiology,* 99**,** 64-74 DOI: https://doi.org/10.1016/j.jclinepi.2018.03.005

OWEN, R. K., TINCELLO, D. G. & KEITH, R. A. 2015. Network Meta-Analysis: Development of a Three-Level Hierarchical Modeling Approach Incorporating Dose-Related Constraints. *Value in Health,* 18**,** 116-126 DOI: http://dx.doi.org/10.1016/j.jval.2014.10.006.

PAGE, M. J., HIGGINS, J. P. T., CLAYTON, G., STERNE, J. A. C., HROBJARTSSON, A. & SAVOVIC, J. 2016. Empirical evidence of study design biases in randomized trials: systematic review of meta-epidemiological studies. *PLoS ONE,* DOI: https://doi.org/10.1371/journal.pone.0159267.

PAPANIKOS, T., THOMPSON, J. R., ABRAMS, K. R., STÄDLER, N., CIANI, O., TAYLOR, R. & BUJKIEWICZ, S. 2020. Bayesian hierarchical meta-analytic methods for modeling surrogate relationships that vary across treatment classes using aggregate data. *Statistics in Medicine,* 39**,** 1103-1124 DOI: https://doi.org/10.1002/sim.8465.

PARMAR, M. K. B., TORRI, V. & STEWART, L. 1998. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Statistics In Medicine,* 17**,** 2815-2834 DOI: https://doi.org/10.1002/(sici)1097-0258(19981230)17:24<2815::aid-sim110>3.0.co;2-8.

PEDDER, H., DIAS, S., BENNETTS, M., BOUCHER, M. & WELTON, N. J. 2019. Modelling time-course relationships with multiple treatments: Model-based network meta-analysis for continuous summary outcomes. *Research Synthesis Methods,* 10**,** 267-286 DOI: 10.1002/jrsm.1351 https://onlinelibrary.wiley.com/doi/full/10.1002/jrsm.1351.

PHILLIPPO, D., ADES, A., BELGER, M., BRNABIC, M., SCHACHT, A., SAURE, D., KADZIOLA, Z. & WELTON, N. 2020. Multilevel Network Meta-Regression for population-adjusted treatment comparisons. *Journal of the Royal Statistical Society A,* Revision Under Review.

PHILLIPPO, D. M. 2019. *Calibration of treatment effects in network meta-analysis using individual patient data.* University of Bristol DOI: https://research-information.bris.ac.uk/files/218211125/David_Phillippo_PhD_Thesis.pdf.

PHILLIPPO, D. M., ADES, A. E., DIAS, S., PALMER, S., ABRAMS, K. R. & WELTON, N. J. 2016. NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submissions to NICE. ScHARR, University of Sheffield. DECISION SUPPORT UNIT, S., UNIVERSITY OF SHEFFIELD DOI: http://nicedsu.org.uk/technical-support-documents/population-adjusted-indirect-comparisons-maic-and-stc/.

PHILLIPPO, D. M., ADES, A. E., DIAS, S., PALMER, S., ABRAMS, K. R. & WELTON, N. J. 2018a. Methods for Population-Adjusted Indirect Comparisons in Health Technology Appraisal. *Med Decis Making,* 38**,** 200-211 DOI: https://doi.org/10.1177/0272989X17725740.

PHILLIPPO, D. M., DIAS, S., ADES, A. E., DIDELEZ, V. & WELTON, N. J. 2018b. Sensitivity of treatment recommendations to bias in network meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 181**,** 843-867 DOI: https://doi.org/10.1111/rssa.12341.

PHILLIPPO, D. M., DIAS, S., WELTON, N. J., CALDWELL, D. C., TASKE, N. & ADES, A. E. 2019. Threshold Analysis as an Alternative to GRADE for Assessing Confidence in Guideline Recommendations Based on Network Meta-analyses. *Annals of Internal Medicine,* 170**,** 538-546 DOI: 10.7326/M18-3542 https://annals.org/aim/article-abstract/2729209/threshold-analysis-alternative-grade-assessing-confidence-guideline-recommendations-based-network.

PREVOST, T. C., ABRAMS, K. R. & JONES, D. R. 2000. Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Statistics in Medicine,* 19**,** 3359-3376 DOI: https://doi.org/10.1002/1097-0258(20001230)19:24<3359::AID-SIM710>3.0.CO;2-N.

PRICE, M. J. & BRIGGS, A. H. 2002. Development of an economic model to assess the cost effectiveness of asthma management strategies. *Pharmacoeconomics,* 20**,** 183-194 DOI: htpps://doi.org/10.2165/00019053-200220030-00004.

PRICE, M. J., WELTON, N. J. & ADES, A. E. 2011. Parameterisation of treatment effects for meta-analysis in multi-state Markov models. *Statistics In Medicine,* 30**,** 140-151 DOI: https://doi.org/10.1002/sim.4059.

PUHAN, M. A., SCHÜNEMANN, H. J., MURAD, M. H., LI, T., BRIGNARDELLO-PETERSEN, R., SINGH, J. A., KESSELS, A. G. & GUYATT, G. H. 2014. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. *BMJ,* 349**,** g5630 DOI: https://doi.org/10.1136/bmj.g5630.

REITSMA, J. B., GLAS, A. S., RUTJES, A. W. S., SCHOLTEN, R. J. P. M., BOSSUYT, P. M. & ZWINDERMAN, A. H. 2005. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology,* 58**,** 982-990 DOI: https://doi.org/10.1016/j.jclinepi.2005.02.022

REN, S., OAKLEY, J. E. & STEVENS, J. W. 2018. Incorporating Genuine Prior Information about Between-Study Heterogeneity in Random Effects Pairwise and Network Meta-analyses. *Medical Decision Making,* 38**,** 531-542 DOI: https://doi.org/10.1177/0272989x18759488

RHODES, K. M., TURNER, R. M. & HIGGINS , J. P. T. 2015. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *Journal of Clinical Epidemiology,* 68**,** 52-60 DOI: https://doi.org/10.1016/j.jclinepi.2014.08.012.

RHODES, K. M., TURNER, R. M., WHITE, I. R., JACKSON, D., SPIEGELHALTER, D. J. & HIGGINS, J. P. T. 2016. Implementing informative priors for heterogeneity in meta-analysis using meta-regression and pseudo data. *Statistics in Medicine,* 35**,** 5495-5511 DOI: https://doi.org/10.1002/sim.7090.

RIEMSMA, R., LHACHIMI, S. K., ARMSTRONG, N., VAN ASSELT, A. D. I., ALLEN, A., MANNING, N., HARKER, J., TUSHABE, D. A., SEVERENS, J. L. & KLEIJNEN, J.

2011. Roflumilast for the management of severe chronic obstructive pulmonary disease: a Single Technology Appraisal. Kleijnen Systematic Reviews Ltd. DOI: https://pdfs.semanticscholar.org/64ab/53d723b2b840a25b635f9a596bbb526caddd.pdf.

RILEY, R. 2009. Multivariate meta-analyis: the effect of ignoring within-study correlation. *Journal of the Royal Statistical Society (A),* 172**,** 789-811 DOI: https://doi.org/10.1136/bmj.j3932.

RILEY, R. D., ABRAMS, K. R., LAMBERT, P. C., SUTTON, A. J. & THOMPSON, J. R. 2007. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics In Medicine,* 26**,** 78-97 DOI: https://doi.org/10.1002/sim.2524.

RILEY, R. D., JACKSON, D., SALANTI, G., BURKE, D. L., PRICE, M., KIRKHAM, J. & WHITE, I. R. 2017. Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. *BMJ,* 358**,** j3932 DOI: https://doi.org/0.1136/bmj.j3932.

RILEY, R. D., PRICE, M. J., JACKSON, D., WARDLE, M., GUEYFFIER, F., WANG, J., STAESSEN, J. A. & WHITE, I. R. 2014. Multivariate meta-analysis using individual participant data. *Research Synthesis Methods,* 6**,** 157-174 DOI: 10.1002/jrsm.1129.

RILEY, R. D., PRICE, M. J., JACKSON, D., WARDLE, M., GUEYFFIER, F., WANG, J., STAESSEN, J. A. & WHITE, I. R. 2015. Multivariate meta-analysis using individual participant data. *Research Synthesis Methods,* 6**,** 157-174 DOI: https://doi.org/10.1002/jrsm.1129.

RILEY, R. D., THOMPSON, J. R. & ABRAMS, K. R. 2008. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics,* 9**,** 172-186 DOI: https://doi.org/10.1093/biostatistics/kxm023.

RIZOPOULOS, D. 2012. *Joint Models for Longitudinal and Time-to-Event Data With Applications in R*, Chapman and Hall/CRC.

RÖVER, C., WANDEL, S. & FRIEDE, T. 2019. Model averaging for robust extrapolation in evidence synthesis. *Statistics in Medicine,* 38**,** 674-694 DOI: https://doi.org/10.1002/sim.7991.

ROYSTON, P. & ALTMAN, D. G. 1994. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* 43**,** 429-467 DOI: https://doi.org/10.2307/2986270.

ROYSTON, P. & PARMAR, M. 2011. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trial when the proportional hazards assumption is in doubt. *Statistics in Medicine,* 30**,** 2409- 2421 DOI: https://doi.org/10.1002/sim.4274.

ROYSTON, P. & PARMAR, M. K. B. 2002. Flexible parameteric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics In Medicine,* 21**,** 2175-2197 DOI: https://doi.org/10.1002/sim.1203.

ROYSTON, P. & PARMAR, M. K. B. 2013. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology,* 13**,** 152-152 DOI: https://doi.org/10.1186/1471-2288-13-152.

RUCKER, G., PETROPOULOU, M. & SCHWARZER, G. 2019. Network meta-analysis of multicomponent interventions. *Biometrical Journal,* DOI: https://doi.org/10.1002/bimj.201800167.

RÜCKER, G., SCHWARZER, G., KRAHN, U. & KÖNIG, J. 2015. netmeta: Network Meta-Analysis using Frequentist Methods. version 0.8-0 ed.

RUTHERFORD, M., LAMBERT, P.C., SWEETING, M.J., PENNINGTON, R., CROWTHER, M.J., ABRAMS, K.R., LATIMER, N.R. 2020. NICE DSU Technical Support Document 21. Flexible Methods for Survival

Analysis. ScHARR, University of Sheffield. DECISION SUPPORT UNIT, S., UNIVERSITY OF SHEFFIELD.

RUTTER, C. M. & GATSONIS, C. A. 2001. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine,* 20**,** 2865-2884 DOI: https://doi.org/10.1002/sim.942.

SALANTI, G., DEL GIOVANE, C., CHAIMANI, A., CALDWELL, D. M. & HIGGINS, J. P. T. 2014. Evaluating the Quality of Evidence from a Network Meta-Analysis. *PLoS ONE,* 9**,** e99682 DOI: https://doi.org/10.1371/journal.pone.0099682.

SALANTI, G., DIAS, S., WELTON, N. J., ADES, A. E., GOLFINOPOULOS, V., KYRGIOU, M., MAURI, D. & IOANNIDIS, J. P. A. 2010. Evaluating novel agent effects in multiple treatments meta-regression. *Statistics In Medicine,* 29**,** 2369-2383 DOI: https://doi.org/10.1002/sim.4001

SARAMAGO, P., CHUANG, L.-H. & SOARES, M. 2014. Network meta-analysis of (individual patient) time to event data alongside (aggregate) count data. *BMC Medical Research Methodology,* 14**,** 105 DOI: https://doi.org/10.1186/1471-2288-14-105.

SAVOVIC, J., JONES, H., ALTMAN, D., HARRIS, R., JUNI, P., PILDAL, J., ALS-NIELSEN, B., BALK, E. M., GLUUD, C., GLUUD, L. L., IOANNIDIS, J. P. A., SCHULZ, K. F., BEYNON, R., WELTON, N. J., WOOD, L., MOHER, D., DEEKS, J. J. & STERNE, J. 2012a. Influence of study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. *Health Technology Assessment,* 16 DOI: https://doi.org/10.3310/hta16350.

SAVOVIC, J., JONES, H. E., ALTMAN, D., HARRIS, R., JUNI, P., PILDAL, J. & AL., E. 2012b. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. *Health Technology Assessment,* 16**,** 1-82 DOI: https://doi.org/10.3310/hta16350.

SAVOVIC, J., JONES, H. E., ALTMAN, D. G., HARRIS, R. J., JUNI, P., PILDAL, J., ALS-NIELSEN, B., BALK, E. M., GLUUD, C., GLUUD, L. L., IOANNIDIS, J. P. A., SCHULZ, K. F., BEYNON, R., WELTON, N. J., WOOD, L., MOHER, D., DEEKS, J. J. & STERNE, J. A. C. 2012c. Influence of Reported Study Design Characteristics on Intervention Effect Estimates From Randomized, Controlled Trials. *Annals of Internal Medicine,* 157**,** 429-438 DOI: https://doi.org/10.7326/0003-4819-157-6-201209180-00537.

SAVOVIĆ, J., TURNER, R. M., MAWDSLEY, D., JONES, H. E., BEYNON, R., HIGGINS, J. & STERNE, J. A. C. 2018. Association between risk-of-bias assessments and results of randomized trials in Cochrane reviews: the ROBES meta-epidemiologic study. *American Journal of Epidemiology,* 187**,** 1113-1122 DOI: https://doi.org/10.1093/aje/kwx344.

SCHILLER, I., VAN SMEDEN, M., HADGU, A., LIBMAN, M., REITSMA, J. B. & DENDUKURI, N. 2016. Bias due to composite reference standards in diagnostic accuracy studies. *Statistics in Medicine,* 35**,** 1454-1470 DOI: https://doi.org/10.1002/sim.6803

SCHMITZ, S., ADAMS, R. & WALSH, C. 2013. Incorporating data from various trial designs into a mixed treatment comparison model. *Statistics in Medicine,* 32**,** 2935-2949 DOI: https//doi.org/10.1002/sim.5764.

SCHNELL-INDERST, P., IGLESIAS, C. P., ARVANDI, M., CIANI, O., MATTEUCCI GOTHE, R., PETERS, J., BLOM, A. W., TAYLOR, R. S. & SIEBERT, U. 2017. A bias-adjusted evidence synthesis of RCT and observational data: the case of total hip replacement. *Health Economics,* 26**,** 46-69 DOI: https://doi.org/10.1002/hec.3474.

SENN, S., GAVINI, D. M. & SCHEEN, A. 2013. Issues in performing a network meta-analysis. *Statistical Methods in Medical Research,* 22**,** 169-189 DOI: https://doi.org/10.1177/0962280211432220.

SIANNIS, F., BARRETT, J., FAREWELL, V. & TIERNEY, J. 2010. One-stage parametric meta-analysis of time-to-event outcomes. *Statistics in Medicine,* 29**,** 3030-3045 DOI: https://doi.org/10.1002/sim.4086.

SIGNOROVITCH, J. E., SIKIRICA, V., ERDER, M. H., XIE, J. P., LU, M., HODGKINS, P. S., BETTS, K. A. & WU, E. Q. 2012. Matching-Adjusted Indirect Comparisons: A New Tool for Timely Comparative Effectiveness Research. *Value in Health,* 15**,** 940-947 DOI: https://doi.org/10.1016/j.jval.2012.05.004.

SIGNOROVITCH, J. E., WU, E. Q., YU, A. P., GERRITS, C. M., KANTOR, E., BAO, Y. J., GUPTA, S. R. & MULANI, P. M. 2010. Comparative Effectiveness Without Head-to-Head Trials A Method for Matching-Adjusted Indirect Comparisons Applied to Psoriasis Treatment with Adalimumab or Etanercept. *Pharmacoeconomics,* 28**,** 935-945 DOI: https://doi.org/10.2165/11538370-000000000-00000.

SOARES, M. O., DUMVILLE, J., ADES, A. E. & WELTON, N. J. 2014. Treatment comparisons for decision making: facing the problems of sparse and few data. . *Journal of the Royal Statistical Society (A),* 177**,** 259-279 DOI: https://doi.org/10.1111/rssa.12010.

SPINELI, L. M. 2019. Modeling missing binary outcome data while preserving transitivity assumption yielded more credible network meta-analysis results. *Journal of Clinical Epidemiology,* 105**,** 19-26 DOI: https://doi.org/10.1016/j.jclinepi.2018.09.002.

SPINELI, L. M., HIGGINS, J. P. T., CIPRIANI, A., LEUCHT, S. & SALANTI, G. 2013. Evaluating the impact of imputations for missing participant outcome data in a network meta-analysis. *Clinical Trials,* 10**,** 378-388 DOI: https://doi.org/10.1177/1740774512470317.

SPINELI, L. M., KALYVAS, C. & PATERAS, K. 2019. Participants' outcomes gone missing within a network of interventions: Bayesian modeling strategies. *Statistics in Medicine,* 38**,** 3861-3879 DOI: https://doi.org/10.1002/sim.8207.

STEINHAUSER, S., SCHUMACHER, M. & RUCKER, G. 2016. Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. *BMC Med Res Methodol,* 16**,** 97 DOI: https://doi.org/10.1186/s12874-016-0196-1

STETTLER, C., ALLEMANN, S., WANDEL, S., KASTRATI, A., MORICE, M. C., SCHOMIG, A., PFISTERER, M. E., STONE, G. W., LEON, M. B., DE LEZO, J. S., GOY, J. J., PARK, S. J., SABATE, M., SUTTORP, M. J., KELBAEK, H., SPAULDING, C., MENICHELLI, M., VERMEERSCH, P., DIRKSEN, M. T., CERVINKA, P., DE CARLO, M., ERGLIS, A., CHECHI, T., ORTOLANI, P., SCHALIJ, M. J., DIEM, P., MEIER, B., WINDECKER, S. & JUNI, P. 2008. Drug eluting and bare metal stents in people with and without diabetes: collaborative network meta-analysis. *British Medical Journal,* 337**,** 1331 DOI: https://doi.org/10.1136/bmj.a1331.

STETTLER, C., WANDEL, S., ALLEMANN, S., KASTRATI, A., MORICE, M. C., SCHOMIG, A., PFISTERER, M. E., STONE, G. W., LEON, M. B., DE LEZO, J. S., GOY, J. J., PARK, S. J., SABATE, M., SUTTORP, M. J., KELBAEK, H., SPAULDING, C., MENICHELLI, M., VERMEERSCH, P., DIRKSEN, M. T., CERVINKA, P., PETRONIO, A. S., NORDMANN, A. J., DIEM, P., MEIER, B., ZWAHLEN, M., REICHENBACH, S., TRELLE, S., WINDECKER, S. & JUNI, P. 2007. Outcomes associated with drug-eluting and bare-metal stents: a collaborative network meta-analysis. *Lancet,* 370**,** 937-948 DOI: https://doi.org/10.1016/S0140-6736(07)61444-5

SULLIVAN, T., LATIMER, N., GRAY, J., SORICH, M., SALTER, A. & KARNON, J. 2020. Adjusting for Treatment Switching in Oncology Trials: A Systematic Review and Recommendations for Reporting. *Value in Health,* 23**,** 388-396 DOI: https://doi.org/10.1016/j.jval.2019.10.015.

SUTTON, A. J. & ABRAMS, K. R. 2001. Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research,* 10**,** 277-303 DOI: https://doi.org/10.1177/096228020101000404.

TAKWOINGI, Y., LEEFLANG, M. M. G. & DEEKS, J. J. 2013. Empirical Evidence of the Importance of Comparative Studies of Diagnostic Test Accuracy. *Annals of Internal Medicine,* 158**,** 544-+ DOI: https://doi.org/10.7326/0003-4819-158-7-201304020-00006

TAN, S. H., ABRAMS, K. R. & BUJKIEWICZ, S. 2018. Bayesian Multiparameter Evidence Synthesis to Inform Decision Making: A Case Study in Metastatic Hormone-Refractory Prostate Cancer. *Medical Decision Making,* 38**,** 834-848 DOI: https://doi.org/10.1177/0272989x18788537.

TAYLOR, R. S. & ELSTON, J. 2009. The use of surrogate outcomes in model-based cost-effectiveness analyses: a survey of UK health technology assessment reports. *Health Technology Assesment,* 13 DOI: https://doi.org/10.3310/hta13080.

THORLUND, K., MILLS, E. J., WU, P., RAMOS, E., CHATTERJEE, A., DRUYTS, E. & GOADSBY, P. J. 2014. Comparative efficacy of triptans for the abortive treatment of migraine: A multiple treatment comparison meta-analysis. *Cephalalgia,* 34**,** 258-267 DOI: https://doi.org/10.1177/0333102413508661.

TRIKALINOS, T. A., HOAGLIN, D. C. & SCHMID, C. H. 2014a. An empirical comparison of univariate and multivariate meta-analyses for categorical outcomes. *Statistics in Medicine,* 33**,** 1441-1459 DOI: https://doi.org/10.1002/sim.6044.

TRIKALINOS, T. A., HOAGLIN, D. C., SMALL, K. M., TERRIN, N. & SCHMID, C. H. 2014b. Methods for the joint meta-analysis of multiple tests. *Res Synth Methods,* 5**,** 294-312 DOI: https://doi.org/10.1002/jrsm.1115.

TURNER, N. L., DIAS, S., ADES, A. E. & WELTON, N. J. 2015a. A Bayesian framework to account for uncertainty due to missing binary outcome data in pairwise meta-analysis. *Statistics in Medicine,* 34**,** 2062-2080 DOI: https://doi.org/10.1002/sim.6475.

TURNER, R. M., DAVEY, J., CLARKE, M. J., THOMPSON, S. G. & HIGGINS, J. P. T. 2012. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *International Journal of Epidemiology,* 41**,** 818-827 DOI: https://doi.org/10.1093/ije/dys041.

TURNER, R. M., DOMÍNGUEZ-ISLAS, C. P., JACKSON, D., RHODES, K. M. & WHITE, I. R. 2019. Incorporating external evidence on between-trial heterogeneity in network meta-analysis. *Statistics in Medicine,* 38**,** 1321-1335 DOI: https://doi.org/10.1002/sim.8044.

TURNER, R. M., JACKSON, D., WEI, Y., THOMPSON, S. G. & HIGGINS, J. P. T. 2015b. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Statistics in Medicine,* 34**,** 984-998 DOI: https://doi.org/10.1002/sim.6381.

TURNER, R. M., SPIEGELHALTER, D. J., SMITH, G. C. S. & THOMPSON, S. G. 2009. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society (A),* 172**,** 21-47 DOI: https://doi.org/10.1111/j.1467-985X.2008.00547.x.

VAN GELOVEN, N., SWANSON, S., RAMSPEK, C., LUIJKEN, K., VAN DIEPEN, M., MORRIS, T., GROENWOLD, R., VAN HOUWELINGEN, H., PUTTER, H. & LE CESSIE, S. 2020. Prediction meets causal inference: the role of treatment in clinical prediction models. *arXiv,* DOI: https://arxiv.org/pdf/2004.06998.pdf.

VERDE, P. E. & OHMANN, C. 2015. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. *Research Synthesis Methods,* 6**,** 45-62 DOI: https://doi.org/10.1002/jrsm.1122.

WALTER, S. D., IRWIG, L. & GLASZIOU, P. P. 1999. Meta-analysis of diagnostic tests with imperfect reference standards. *Journal of Clinical Epidemiology,* 52**,** 943-951 DOI: https://doi.org/10.1016/S0895-4356(99)00086-4

WARREN, F. C., ABRAMS, K. R. & SUTTON, A. J. 2014. Hierarchical network meta-analysis models to address sparsity of events and differing treatment classifications with regard to adverse outcomes. *Statistics in Medicine,* 33**,** 2449-2466 DOI: https://doi.org/10.1002/sim.6131.

WEI, Y. & HIGGINS , J. 2013a. Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. *Statistics In Medicine,* 32**,** 1191-1205 DOI: https://doi.org/10.1002/sim.5679.

WEI, Y. & HIGGINS , J. P. T. 2013b. Bayesian multivariate meta-analysis with multiple outcomes. *Statistics in Medicine,* 32**,** 2911-2934 DOI: https://doi.org/10.1002/sim.5745.

WEI, Y. & ROYSTON, P. 2017. Reconstructing time-to-event data from published Kaplan–Meier curves. . *The Stata Journal,* 17**,** 786–802 DOI: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5796634/.

WEI, Y., ROYSTON, P., TIERNEY, J. & PARMAR, M. 2015. Meta-analysis of time-to-event outcomes from randomized trials using restricted mean survival time: application to individual participant data. *Statistics in Medicine,* 34**,** 2881- 2898 DOI: https://doi.org/10.1002/sim.6556

WELTON, N. J. & ADES, A. E. 2005. Estimation of Markov Chain Transition Probabilities and Rates from Fully and Partially Observed Data: Uncertainty Propagation, Evidence Synthesis and Model Calibration. *Medical Decision Making,* 25**,** 633-645 DOI: https://doi.org/10.1177/0272989X05282637

WELTON, N. J., ADES, A. E., CARLIN, J. B., ALTMAN, D. G. & STERNE, J. A. C. 2009a. Models for potentially biased evidence in meta-analysis using empirically based priors. *Journal of the Royal Statistical Society (A),* 172**,** 119-136 DOI: https://doi.org/10.1111/j.1467-985X.2008.00548.x.

WELTON, N. J., CALDWELL, D. M., ADAMOPOULOS, E. & VEDHARA, K. 2009b. Mixed Treatment Comparison Meta-analysis of Complex Interventions: Psychological interventions in coronary heart disease. *American Journal Of Epidemiology,* 169**,** 1158-1165 DOI: https://doi.org/10.1093/aje/kwp014.

WELTON, N. J., COOPER, N. J., ADES, A. E., LU, G. & SUTTON, A. J. 2008. Mixed treatment comparison with multiple outcomes reported inconsistently across trials: evaluation of antivirals for treatment of influenza A and B. *Statistics In Medicine,* 27**,** 5620-5639 DOI: https://doi.org/10.1002/sim.3377.

WELTON, N. J., WILLIS, S. R. & ADES, A. E. 2010. Synthesis of Survival and Disease Progression Outcomes for Health Technology Assessment of Cancer Therapies. *Research Synthesis Methods,* 1**,** 239-257 DOI: https://doi.org/10.1002/jrsm.21.

WHITE, I., HIGGINS , J. & WOOD, A. M. 2008a. Allowing for uncertainty due to missing data in meta-analysis - Part 1: Two-stage methods. *Statistics In Medicine,* 27**,** 711-727 DOI: https://doi.org/10.1002/sim.3008.

WHITE, I., TURNER, R., KARAHALIOS, A. & SALANTI, G. 2019. A comparison of arm-based and contrast-based models for network meta-analysis. *Statistics in Medicine,* 38**,** 5197-5213 DOI: https://doi.org/10.1002/sim.8360.

WHITE, I. R. 2015. Network meta-analysis. *Stata Journal,* 15**,** 951-985 DOI: https://www.stata-journal.com/article.html?article=st0410.

WHITE, I. R., WOOD, A., WELTON, N. J., ADES, A. E. & HIGGINS, J. P. T. 2008b. Allowing for uncertainty due to missing data in meta-analysis - Part 2: Hierarchical models. *Statistics In Medicine,* 27**,** 728-745 DOI: https://doi.org/10.1002/sim.3007.

WILLIAMS, C., LEWSEY, J., MACKAY, D. & BRIGGS, A. 2017. Estimation of Survival Probabilities for Use in Cost-effectiveness Analyses: A Comparison of a Multi-state Modeling Survival Analysis Approach with Partitioned Survival and Markov Decision-Analytic Modeling. *Medical Decision Making,* 37**,** 427-439 DOI: https://doi.org/10.1177/0272989X16670617.

WILLIAMSON, P. R., SMITH, C. T., HUTTON, J. L. & MARSON, A. G. 2002. Aggregate data meta-analysis with time-to-event outcomes. *Statistics In Medicine,* 21**,** 3337-3351 DOI: https://doi.org/10.1002/sim.1303.

WOODS, B., SIDERIS, E., PALMER, S., LATIMER, N. & SOARES, M. 2017. NICE DSU Technical Support Document 19: Partitioned survival analysis for decision modelling in health care: a critical review. DOI: http://nicedsu.org.uk/wp-content/uploads/2017/06/Partitioned-Survival-Analysis-final-report.pdf.

YUAN, Y. & LITTLE, R. J. A. 2009. Meta-Analysis of Studies with Missing Data. *Biometrics,* 65**,** 487-496 DOI: https://doi.org/10.1111/j.1541-0420.2008.01068.x.

ZHANG, J., CARLIN, B. P., NEATON, J. D., SOON, G. G., NIE, L., KANE, R., VIRNIG, B. A. & CHU, H. 2014. Network meta-analysis of randomized clinical trials: reporting the proper summaries. *Clinical Trials,* 11**,** 246-262 DOI: https://doi.org/10.1177/1740774513498322.

ZHENG, Y., PAN, F. & SORENSEN, S. 2017. Modeling Treatment Sequences in Pharmacoeconomic Models. *Pharmacoeconomics,* 35**,** 15-24 DOI: https://doi.org/10.1007/s40273-016-0455-3.