NICE DSU TECHNICAL SUPPORT DOCUMENT 18: METHODS FOR POPULATION-ADJUSTED INDIRECT COMPARISONS IN SUBMISSIONS TO NICE

REPORT BY THE DECISION SUPPORT UNIT

December 2016

David M. Phillippo,[1] A. E. Ades,[1] Sofia Dias,[1]
Stephen Palmer,[2] Keith R. Abrams,[3] Nicky J. Welton[1]

[1] School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, UK

[2] Centre for Health Economics, University of York

[3] Department of Health Sciences, University of Leicester

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) is a collaboration between the Universities of Sheffield, York and Leicester. We also have members at the University of Bristol, London School of Hygiene and Tropical Medicine and Brunel University. 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 Technology Appraisal Programme. Please see our website for further information www.nicedsu.org.uk

ABOUT THE TECHNICAL SUPPORT DOCUMENT SERIES

The NICE Guide to the Methods of Technology Appraisal[i] is a regularly updated document that provides an overview of the key principles and methods of health technology assessment and appraisal for use in NICE appraisals. The Methods Guide does not provide detailed advice on how to implement and apply the methods it describes. This DSU series of Technical Support Documents (TSDs) is intended to complement the Methods Guide by providing detailed information on how to implement specific methods.

The TSDs provide a review of the current state of the art in each topic area, and make clear recommendations on the implementation of methods and reporting standards where it is appropriate to do so. They aim to provide assistance to all those involved in submitting or critiquing evidence as part of NICE Technology Appraisals, whether companies, assessment groups or any other stakeholder type. We recognise that there are areas of uncertainty, controversy and rapid development. It is our intention that such areas are indicated in the TSDs. All TSDs are extensively peer reviewed prior to publication (the names of peer reviewers appear in the acknowledgements for each document). Nevertheless, the responsibility for each TSD lies with the authors and we welcome any constructive feedback on the content or suggestions for further guides.

Please be aware that whilst the DSU is funded by NICE, these documents do not constitute formal NICE guidance or policy.

Professor Allan Wailoo
Director of DSU and TSD series editor.

---

[i] National Institute for Health and Care Excellence. Guide to the methods of technology appraisal, 2013 (updated April 2013), London

Acknowledgements

The authors wish to acknowledge the contributions of Richard Grieve, Jeroen Jansen, Andreas Karabis, James Signorovitch, Ian White and the NICE team led by Rosie Lovett, who provided peer comments on the draft document.

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.

This report should be referenced as follows:

Phillippo, D.M., Ades, A.E., Dias, S., Palmer, S., Abrams, K.R., Welton, N.J. NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submission to NICE. 2016. Available from http://www.nicedsu.org.uk

EXECUTIVE SUMMARY

This Technical Support Document examines methods for *population-adjusted indirect comparisons*, in which individual patient data (IPD) in one or more trials are used to adjust for between-trial differences in the distribution of variables that influence outcome. Recently proposed methods are the *Matching-Adjusted Indirect Comparison* (MAIC) and the *Simulated Treatment Comparison* (STC). We look at the theory behind MAIC and STC methods, review published applications, briefly touch on alternative methods, and make recommendations on the use of population-adjusted estimates in submissions to NICE.

Methods that "map" the treatment effects observed in one population into effects that would be observed in another population have existed for many years, coming under the general heading of "standardisation". These methods have been applied to treatment effects from both randomised and non-randomised studies. We briefly review the literature on *propensity score weighting* and on *outcome regression* methods, which underlie MAIC and STC respectively, as well as doubly robust estimation which combines features of both approaches (Section 2.1). The properties of these methods have also been examined in the related literature on population average treatment effects.

The novelty of MAIC and STC is that they apply the classic propensity score and regression methods to the specific case of indirect comparisons, with limited availability of IPD. Application of population adjustment methods to indirect comparisons is well-motivated. Standard methods for indirect comparisons and network meta-analysis, proposed by Bucher *et al.*[1] and then Dias *et al.*[2], are based on aggregate data. The key assumption behind these standard methods is that there is no difference between the trials in the distribution of effect-modifying variables. Given the high levels of heterogeneity often found in trial networks, as evidenced by the frequent use of "random effects" models, the validity of indirect comparisons in the very sparse networks seen in many submissions to NICE must be considered carefully. In networks consisting of only one or two trials per treatment, indirect comparisons are highly vulnerable to systematic variation (bias) resulting from imbalances in effect modifier distributions. Under these circumstances population-adjusted methods have a distinct attraction.

Drawbacks of MAIC and STC as methods for population adjustment

Despite the motivation for population adjustment methods, the actual form of population adjustment typically implemented by MAIC and STC suffers from several very considerable drawbacks. These are outlined in our analysis of MAIC and STC methodology (Section 2), and examples are given in our review of published MAIC and STC applications (Section 3). Principally:

1. Both MAIC and STC can be used to carry out either an "anchored" indirect comparison, where there is a common comparator arm in each trial, or an "unanchored" indirect comparison, where

there is a disconnected treatment network or single-arm studies. An unanchored MAIC or STC effectively assumes that absolute outcomes can be predicted from the covariates; that is, it assumes that all effect modifiers and prognostic factors are accounted for. This assumption is very strong, and largely considered impossible to meet. Failure of this assumption leads to an unknown amount of bias in the unanchored estimate. We have suggested possible methods for providing a plausible range for the residual bias via estimation of the unexplained heterogeneity in absolute outcomes (Appendix C).

2. In most applications of MAIC and STC (9 out of 11 published applications to date identified in our review), little evidence is presented that population adjustment will produce more accurate estimates than standard methods for indirect comparisons. Specifically, in the large majority of cases no evidence is presented prior to analysis that any given covariate is an effect modifier, or that the degree of imbalance is sufficient for adjustment to make a material difference.
3. While standard methods for indirect comparisons assume additivity of effects on a pre-specified transformed scale, such as the logit scale for probabilities, MAIC and STC typically assume additivity on the natural outcome scale. This represents a major departure from the way in which relative treatment effects are usually conceptualised in health technology assessment (HTA).
4. MAIC and STC have been designed to meet a very specific situation applying to companies making submissions to NICE, in which companies have access to individual patient data (IPD) from their own trials, say on treatments A and B, labelled as AB, but only aggregate outcomes (as summarised in publication, for example as means and confidence intervals for outcomes in each arm) and marginal covariate information (for example proportion of females, mean and standard deviation of age) from a competitor's trials, say on treatments A and C, labelled as AC. By making certain assumptions about the joint distribution of covariates in the AC trial(s), MAIC and STC set out to generate the AB effect that would be observed in the AC trial population. Thus, companies deploying MAIC or STC are not only arguing that the treatment effect is dependent on the population, but they are further assuming that the target population is closer to that represented in the competitor trial than in their own trial. In reality, the target population for a decision is likely to be represented by a UK cohort or registry study, and may differ from both AB and AC trials.

Population-adjusted estimates of treatment effects in the context of submissions to NICE

While there is a clear rationale for considering population-adjusted estimates of treatment effects, there is a lack of clarity about exactly how, and when, they should be applied in practice, and even whether the results are relevant to the decision problem. This increases the risk that assumptions being made in one submission are fundamentally different from – even incompatible with – the assumptions being made a year later in another on the same condition. In the interests of transparency, and to ensure a

degree of certainty for those making submissions, the recommendations below attempt to regularise how and when population-adjustment should be used, and set out additional analyses that should be presented to support their use and assist their interpretation. These recommendations can be no more than provisional, as comprehensive simulation studies are required to explore the properties of the various methods available. In Appendix A we provide flow charts which summarise how methods for indirect comparison should be chosen, how they are implemented, and how their results should be presented. In Appendix D we provide a worked example of MAIC and STC carried out according to our recommendations, complete with accompanying R code.

Summary of recommendations

1. When connected evidence with a common comparator is available, only "anchored" forms of population adjustment may be used. "Unanchored" population adjustment may only be considered in the absence of a connected network of randomised studies, or where there are single-arm studies involved. (Section 4.2.2.)
2. Submissions using anchored population adjustment must produce evidence that population adjustment is likely to produce less biased estimates than would be available through standard indirect comparisons. This requires (i) showing there are grounds for believing one or more of the available covariates is an effect modifier, and (ii) showing that there is sufficient imbalance in those effect modifiers to result in a material bias, in relation to the observed relative treatment effect. (Section 4.2.3.)
3. Submissions using unanchored forms of population adjustment must provide evidence on the likely extent of error due to unaccounted for covariates, in relation to the observed relative treatment effect. (Section 4.2.4.)
4. For anchored indirect comparisons performed via propensity score weighting methods (e.g. MAIC), all effect modifiers should be adjusted for to ensure balance and reduce bias, but no purely prognostic variables to avoid inflating standard error due to over-matching. For anchored indirect comparisons performed via outcome regression methods (e.g. STC), all effect modifiers in imbalance should be adjusted for to reduce bias, and further effect modifiers and prognostic variables may be adjusted for if this improves model fit to reduce standard error. For an unanchored indirect comparison, population adjustment methods should adjust for all effect modifiers and prognostic variables. (Section 4.2.5.)
5. Indirect comparisons must be carried out on the usual linear predictor scale used for evidence synthesis of that outcome. (Section 4.2.6.)
6. The target population for the decision problem must be explicitly stated, and the population adjustment must deliver treatment effect estimates for that target population. (Section 4.2.7.)

7. Strict reporting requirements are recommended, including the assessment of covariate distributions, evidence for effect modifier status, distribution of weights (if applicable), and appropriate measures of uncertainty. (Section 4.2.8.)

Overall conclusions

There is a clear role for population-adjusted indirect comparisons in Health Technology Assessment, although their use in each case must be justified. Unanchored methods for population adjustment are problematic and should not be used when anchored methods can be applied. We propose that population-adjusted indirect comparisons should be carried out on the same scale (log, logit, etc.) that would be used in a standard indirect comparison. Further, population-adjusted estimates of absolute and relative treatment effects can, and should, be constructed specifically for the target decision population. We show how this can be done algebraically and in a worked example.

MAIC and STC and the versions of MAIC and STC that we recommend represent a class of methods that use IPD from one or more studies to predict a population-average outcome for one or more treatments in a different population, and they then effect indirect comparisons at the population level. Other approaches to estimating population-adjusted estimates of absolute and relative treatment effects are also available.

Further research is needed:

- to develop further methods for population adjusted indirect comparisons;
- to assess their comparative vulnerability to failures in assumptions, through comprehensive simulation studies;
- to find ways of estimating the systematic error due to unaccounted covariates in anchored, and especially unanchored, comparisons;
- to investigate the extent of error following from the availability of only marginal, rather than joint, covariate distributions, and to obtain empirical data on the between-trial variation in the joint covariate distributions;
- to extend the methods to larger networks;
- to ensure appropriate uncertainty propagation in population-adjusted estimates; and
- to prepare suitable software tools for population-adjustment with a range of worked examples.

The recommendations made by this TSD should be amended and/or extended in light of subsequent results.

CONTENTS

TABLES AND FIGURES

GLOSSARY OF TERMS

Definitions, where stated here, are not intended to be fully rigorous in the mathematical sense, rather the aim is that they are accessible to the reader. More precise definitions may be found within the referenced literature.

Additive – The effect of a treatment or covariate is said to be additive on a certain scale if the effect of receiving treatment or having a certain covariate value numerically adds or subtracts from a reference value on that scale (as opposed to multiplying or dividing the reference value). The linear predictor scale is by definition additive.

Anchored indirect comparison – An indirect comparison between two treatments which relies on the presence of a common comparator, respecting randomisation within studies to remove bias due to imbalanced prognostic variables. An anchored indirect comparison between treatments B and C in a population P based upon an AB trial and an AC trial is of the form $\Delta_{BC(P)} = \Delta_{AC(P)} - \Delta_{AB(P)}$ $= \left(g\left(Y_{C(P)}\right) - g\left(Y_{A(P)}\right)\right) - \left(g\left(Y_{B(P)}\right) - g\left(Y_{A(P)}\right)\right)$. Does not require as strong assumptions as an unanchored indirect comparison.

Boundedness – An estimator $\hat{\theta}$ for a quantity θ has the property of boundedness if the estimated values always lie within the support of θ; e.g. estimates of probabilities lie between 0 and 1, or estimates of rates are always non-negative.

Confounder – A covariate that is associated with both treatment assignment and outcome (but is not an intermediate variable), such that the treatment effect cannot be disentangled from the effect of the confounders without suitable adjustment.

Consistent estimator – An estimator $\hat{\theta}$ for a quantity θ is consistent if, as the sample size increases to infinity, $\hat{\theta}$ gets ever closer to θ.

Doubly robust – A doubly robust estimator involves specifying two models, one for outcomes and another for sample selection, which are combined in such a way that – as long as at least one model is correct – the estimator is consistent and unbiased. The primary advantage here is that the analyst has two chances to correctly specify a model, compared to using a single method alone.

Effect modifier – A covariate that alters the effect of treatment on outcomes, so that the treatment is more or less effective in different subgroups formed by levels of the effect modifier. Effect modifiers are not necessarily also prognostic variables. Effect modifier status is specific to a given scale: the positive status of a covariate as an effect modifier on one scale does not necessarily imply either positively or negatively effect modifier status on another scale; however, a covariate that is not an effect modifier on one scale is guaranteed to be an effect modifier on another.

Effective Sample Size (ESS) – When estimates are made by weighting a sample, the effective sample size is the number of independent non-weighted individuals that would be required to give an estimate with the same precision as the weighted sample estimate. Weighting always reduces the effective sample size.

Ignorability – A variable T (usually treatment or sample assignment) is ignorable given covariates $\boldsymbol{X}$ if potential outcomes Y are independent of T given $\boldsymbol{X}$. Strong ignorability requires further that every value of T is possible (has probability not 0 or 1) given $\boldsymbol{X}$. For example, strongly ignorable treatment assignment means that there are no unmeasured prognostic variables or effect modifiers in imbalance between the treatment groups, and any individual with given covariate values is not excluded from nor guaranteed either treatment or control.

Linear predictor scale – The scale on which the effects of treatment, effect modification, and prognostic variables are assumed to be additive (that is, linear). For example, the log odds ratio scale is typically used as the linear predictor scale for binary outcomes.

Matching-Adjusted Indirect Comparison (MAIC) – A form of propensity score weighting, applicable where IPD are available in one population and aggregate data in another. Individuals in the IPD population are weighted by the inverse of their propensity score, to balance the covariate distribution with that of target aggregate population. A novel approach to estimating the propensity score must be taken, due to IPD only being available in one of the two populations.

Natural outcome scale – The scale on which an outcome is defined and observed. Typically a link function is used which transforms the data on the natural outcome scale to be modelled on the linear predictor scale. For example, the natural outcome scale for a binary outcome is the probability scale, and a logistic link function is typically used to define a model on the log odds ratio linear predictor scale (this is logistic regression).

Outcome regression – A method for adjusting the outcomes observed in a sample population to those that would have been seen in a target population with a different covariate distribution. A statistical

model is created to describe the outcome in terms of the covariates, and then applied to predict outcomes for the target population.

PATE – Population Average Treatment Effect; the treatment effect that would be observed if the entire population was given the treatment compared to the control.

Prognostic variable – A covariate that affects (or is prognostic of) outcome. We make the distinction between prognostic variables and effect modifiers; effect modifiers are not necessarily also prognostic variables.

Propensity score – The conditional probability of an individual being sampled into a trial, given their covariate values. Propensity scores are typically estimated via logistic regression.

Propensity score weighting – A method for removing differences in the distribution of covariates between two populations (typically one a sample and the other a target population), based on the propensity score of individuals. Individuals in the sample population are weighted by the inverse of their propensity score to balance differences in the covariate distributions.

RCT – Randomised Controlled Trial.

Sandwich estimator – A form of variance estimator that does not rely upon strong assumptions about the data (or in the case of MAIC, the weights), but instead is derived empirically from the data. "Sandwich" refers to how the estimator is constructed, with the empirical approximation "sandwiched" between other matrices.

SATE – Sample Average Treatment Effect; the treatment effect that would be observed in a sample if the entire sample was given the treatment compared to the control. This is the quantity estimated by a RCT, and is equivalent to the PATE when the sample is representative of the population.

Simulated Treatment Comparison – A form of outcome regression, applicable where IPD are available in one population and aggregate data in another. A statistical model describing the outcomes in terms of the covariates is fitted in the IPD population, and used to predict the outcomes that would have been observed in the aggregate target population. The lack of IPD in both populations can lead to numerical methods being used at the prediction stage (hence "simulated").

Unanchored indirect comparison – An indirect comparison between two treatments which does not rely on the presence of a common comparator, and does not respect any randomisation within studies

(if available). An unanchored indirect comparison between treatments B and C in a population P is of the form $\Delta_{BC(P)} = g\left(Y_{C(P)}\right) - g\left(Y_{B(P)}\right)$. Requires much stronger assumptions than an anchored indirect comparison.

Unbiased estimator – An estimator $\hat{\theta}$ for a quantity θ is unbiased if it is correct in expectation, that is $\mathbb{E}\left(\hat{\theta} - \theta\right) = 0$.

1. INTRODUCTION

This Technical Support Document examines methods for *population-adjusted indirect comparisons*, in which individual patient data (IPD) in one or more trials are used to adjust for between-trial differences in the distribution of variables that influence outcome. Recently proposed methods are the *Matching-Adjusted Indirect Comparison* (MAIC)[3-5] and the *Simulated Treatment Comparison* (STC).[4, 6] Other methods are based on network meta-regression with combined IPD and aggregate data.[7-10] MAIC and STC are predicated on the belief that they can "adjust for" between-trial differences in "baseline characteristics", and hence provide a valid estimate of relative treatment effects when standard indirect comparisons are either inappropriate or infeasible. The standard methods for indirect comparisons and network meta-analysis, proposed by Bucher *et al.*[1] and then Dias *et al.*,[2] are based on aggregate data, and make the assumption that the distribution of effect-modifying variables does not differ between studies (see TSDs 1-7[11-17]). A very common scenario arising in submissions to NICE is that where the company has IPD on its own trial, but not on the competitor's trial. There are two ways in which MAIC and STC may be used: (i) in an "anchored" fashion, where connected evidence is available and common comparator arms are taken account of, or (ii) in an "unanchored" fashion, where there is no connected evidence, or comparisons involve single-arm studies.

Population adjustment methods such as MAIC and STC can account for between-trials imbalances in observed covariates. They are not capable of adjusting for differences in, for example, treatment administration, co-treatments, or treatment switching. Between-trials differences of this type are perfectly confounded with treatment. They may be adjusted for at the same time, but using other methods.

The number of submissions to NICE based on MAIC and/or STC have grown since their publication. However, little is known about the reliability or the general properties of these methods, particularly in the context of NICE technology appraisals.

1.1 SCOPE OF THIS TSD

The scope of this document is limited to situations where IPD are available on one or more trials, but only aggregate data (on outcomes and covariates) are available on others. Our work complements TSD 17[18] which focusses on deriving relative treatment effect estimates from non-randomised comparative studies in which IPD are available. There is an area of overlap where TSD 17 considers comparisons based on separate one-arm studies, but where IPD are available from all studies. Despite these differences, our analysis and recommendations dovetail with those of TSD 17. It will be seen that there are parallel themes between the two documents, for example in the families of methods discussed, the assumptions required, and the reliance upon sufficient overlap between population distributions.

In the next section we give a more formal statement of the problem that MAIC and STC set out to solve, and raise a series of issues which will be taken up in the remainder of this report.

1.2 OVERVIEW OF THE PROBLEM

We begin by focussing our attention on the scenario where a connected treatment network is available; the unconnected scenario then follows simply. Consider one AB trial, for which the company has IPD, and one AC trial, for which only published aggregate data are available. We wish to estimate a comparison of the effects of treatments B and C on an appropriate scale in some target population P, denoted by the parameter $d_{BC(P)}$. Throughout this text we make use of bracketed subscripts to denote a specific population. Within the AB population there are parameters $\mu_{A(AB)}$, $\mu_{B(AB)}$ and $\mu_{C(AB)}$ representing the expected outcome on each treatment (including parameters for treatments not studied in the AB trial, e.g. treatment C). The AB trial provides estimators $\bar{Y}_{A(AB)}$ and $\bar{Y}_{B(AB)}$ of $\mu_{A(AB)}$ and $\mu_{B(AB)}$ respectively, which are the summary outcomes, for example the probability of success, on each arm ($\mu_{C(AB)}$ is not estimated by the AB trial). There is a parallel system of parameters ($\mu_{A(AC)}, \mu_{B(AC)}, \mu_{C(AC)}$) and estimators ($\bar{Y}_{A(AC)}, \bar{Y}_{C(AC)}$) in the AC trial.

Having selected a suitable scale, for example a logit, log, risk difference, or mean difference scale, we form estimators $\hat{\Delta}_{AB(AB)}$ and $\hat{\Delta}_{AC(AC)}$ of the trial level (or marginal) relative treatment effects $d_{AB(AB)}$ and $d_{AC(AC)}$ in each trial using the appropriate link function $g(\cdot)$:

$$\hat{\Delta}_{AB(AB)} = g\left(\bar{Y}_{B(AB)}\right) - g\left(\bar{Y}_{A(AB)}\right), \qquad \hat{\Delta}_{AC(AC)} = g\left(\bar{Y}_{C(AC)}\right) - g\left(\bar{Y}_{A(AC)}\right). \tag{1}$$

For example, if a logit scale is selected then the link function is $g(y) = \log\left(y/(1-y)\right)$; for a log scale, the link function is $g(y) = \log(y)$. If the suitable scale is that of the outcome (so no transformation is required), then the identity link is $g(y) = y$.

Throughout this text we make a clear and necessary distinction between *prognostic variables* and *effect modifiers*: prognostic variables are covariates that affect (or are prognostic of) outcome; effect modifiers are covariates that alter the effect of treatment as measured on a given scale, so that treatment is more or less effective depending on the level of the effect modifier. Effect modifiers are not necessarily also prognostic variables, and will be specific to each treatment. Effect modifier status on one scale does not necessarily imply (either positively or negatively) effect modifier status on another scale.

Standard methods for indirect comparison make the assumption that there is no difference in the distribution of trial-level effect modifiers, specific to the chosen scale, between the populations in the AB and AC trials or the target population P, so that marginal relative treatment effects are equal across populations: $d_{AB(AB)} = d_{AB(AC)} = d_{AB(P)}$ and $d_{AC(AB)} = d_{AC(AC)} = d_{AC(P)}$. Under this assumption the standard indirect comparison estimator of the relative effect $d_{BC(P)}$ is

$$\hat{\Delta}_{BC(P)} = \hat{\Delta}_{AC(AC)} - \hat{\Delta}_{AB(AB)}, \tag{2}$$

which takes account of the fact that patients are only randomized *within* trials.

The final step is to apply these relative effects to a specified target population P in which the summary absolute effect (such as the mean change from baseline, or probability of response) of treatment A is $\bar{Y}_{A(P)}$. We can now estimate the summary absolute effects of treatments A, B, C in the target population, $\mu_{A(P)}, \mu_{B(P)}, \mu_{C(P)}$, which have estimators

$$\bar{Y}_{A(P)}, \qquad \hat{Y}_{B(P)} = g^{-1}\left(g\left(\bar{Y}_{A(P)}\right) + \hat{\Delta}_{AB(P)}\right), \qquad \hat{Y}_{C(P)} = g^{-1}\left(g\left(\bar{Y}_{A(P)}\right) + \hat{\Delta}_{AC(P)}\right). \tag{3}$$

(Here we differentiate between: estimators which arise as statistics within the population, denoted with a bar, in this case $\bar{Y}_{A(P)}$ is observed in the P population, and; estimators which are in some sense derived from additional external information and/or assumptions, denoted with a hat, in this case $\hat{Y}_{B(P)}$ and $\hat{Y}_{C(P)}$ require the transfer of information on $d_{AB(AB)}$ and $d_{AC(AC)}$ from the AB and AC populations under the assumption that $d_{AB(P)} = d_{AB(AB)}$ and $d_{AC(P)} = d_{AC(AC)}$.)

Between trial differences in the distribution of prognostic variables (variables related to the outcome) that are *not* effect modifiers do not affect inference, because, as a result of the within-trial randomization, these variables do not impact on the relative treatment effects (assuming that the sample size is sufficiently large and that proper randomisation occurred). Note that effect modifiers X are assumed to have a linear effect on the transformed scale, such that, at any given value of X, the *conditional* relative effect is $d_{AB}(X) = d_{AB}(0) + \gamma X$, conceptualised as an "intercept" term (the relative effect $d_{AB}(0)$ at $X = 0$) plus an interaction effect.

If there are effect modifiers *and if* these are distributed differently between the populations, the relative treatment effects $d_{AB(AB)}, d_{AC(AC)}$ that can be estimated directly from each trial are only valid for a

population with the distribution of effect modifiers observed in that trial. For example, we would have estimates $\hat{d}_{AB(AB)}, \hat{d}_{AC(AC)}$, but it would not be possible to identify a coherent set of estimates, *either* for the population represented in the AB trial $\hat{d}_{AB(AB)}$, $\hat{d}_{AC(AB)}$, $\hat{d}_{BC(AB)} = \hat{d}_{AC(AB)} - \hat{d}_{AB(AB)}$, *or* for the population represented in the AC trial $\hat{d}_{AB(AC)}$, $\hat{d}_{AC(AC)}$, $\hat{d}_{BC(AC)} = \hat{d}_{AC(AC)} - \hat{d}_{AB(AC)}$, or, indeed, in any other target population. It should be noted that adjusting for effect modifiers in a decision-making context is not merely a case of reducing bias in an indirect comparison; the existence of an effect modifier can change the nature of the decision problem: for example if age is considered to be an effect modifier, it raises the possibility that a treatment that is effective at one age might not be effective at another.

The premise of MAIC and STC is to "adjust for" between-trial differences in "baseline characteristics", in order to identify a coherent set of estimates where standard methods of indirect comparison cannot. We describe these methods in detail in Section 2.2.

Briefly, IPD on the AB trial is used to form predictors $\hat{Y}_{A(AC)}, \hat{Y}_{B(AC)}$ of the summary outcomes that would be observed on treatments A and B in the AC trial if the AB trial population was the same as the AC trial population. These predicted outcomes can be based on a regression of the outcome against covariates in the AB trial using IPD, with the regression coefficients then applied to the covariate distribution in the AC trial (STC), or they may be based on a propensity score weighting or matching approach, aimed at reweighting the individuals in the AB trial so that the covariate distribution matches that of the AC trial (MAIC).

The predicted outcomes $\hat{Y}_{A(AC)}, \hat{Y}_{B(AC)}$ may then be used in two ways: relative effects may be estimated by a form of "adjusted" indirect comparison, to use the terminology originated by Glenny *et al.*[19]:

$$\hat{\Delta}_{BC(AC)} = g\left(\overline{Y}_{C(AC)}\right) - g\left(\overline{Y}_{A(AC)}\right) - \left(g\left(\hat{Y}_{B(AC)}\right) - g\left(\hat{Y}_{A(AC)}\right)\right). \quad (4)$$

The term "anchored" is also used in the MAIC/STC literature, and this may be more appropriate because for MAIC/STC the adjusted comparison in (4) typically takes place on the natural outcome scale (see below). Alternatively, an "unadjusted" or "unanchored" indirect comparison can be generated:[4, 5]

$$\hat{\Delta}_{BC(AC)} = g\left(\overline{Y}_{C(AC)}\right) - g\left(\hat{Y}_{B(AC)}\right). \quad (5)$$

This creates a degree of ambiguity about how MAIC/STC is supposed to be used. Furthermore, the literature on MAIC and STC frequently presents indirect comparisons directly on the natural outcome

scale, i.e. with $g(\cdot)$ the identity function in (4) and (5) above, even when a transformed scale is commonly used for standard indirect comparisons and meta-analysis. Using the anchored version, the required estimators of the summary outcomes on treatments A, B, C on the natural outcome scale are $\bar{Y}_{A(AC)}$, $\hat{Y}_{B(AC)} - \hat{Y}_{A(AC)} + \bar{Y}_{A(AC)}$, $\bar{Y}_{C(AC)}$, while in the unanchored version they are $\bar{Y}_{A(AC)}$, $\hat{Y}_{B(AC)}$, $\bar{Y}_{C(AC)}$. The alternative approach, which would be more consistent with the way in which indirect comparisons are usually carried out,[1, 2, 14] would be to carry out the indirect comparison on the transformed scale, obtaining the estimators $\bar{Y}_{A(AC)}$, $g^{-1}\left(g\left(\hat{Y}_{B(AC)}\right) - g\left(\hat{Y}_{A(AC)}\right) + g\left(\bar{Y}_{A(AC)}\right)\right)$, $\bar{Y}_{C(AC)}$.

The potential advantage of methods such as MAIC and STC is that it is only necessary to have IPD on one trial, say the company's AB trial. Specifically, there is no need to have outcome data at the IPD level in the competitor's AC trial, only the aggregate outcomes $\bar{Y}_A, \bar{Y}_C$ alongside the covariate distribution in the AC trial. Note, however, that when IPD are available on the AB trial, MAIC and STC set out to generate a "fair comparison" of treatments A, B, C that is *specific to the population in the AC trial*. The issue of whether the AC population is in fact the target population for decision makers, *rather than the population sampled in the company's own AB trial*, or some other population represented by a UK cohort or register, is therefore a critical issue.

There is a considerable literature on model-based standardisation based on regression adjustment and reweighting, from which MAIC and STC derive. Like MAIC and STC these methods have been aimed at mapping the absolute and relative effects observed in one population into effects that would be predicted in another, in both randomized and observational study settings. The novel aspect of MAIC is to apply specific versions of these methods in the context of indirect comparisons, whilst only having limited access to IPD.

1.3 OBJECTIVES OF THIS REPORT

Based on the above characterization of the problem that MAIC and STC seek to address, this TSD will focus on three key issues:

1. Do MAIC or STC succeed in creating a valid comparison of treatments A, B, C where standard indirect comparisons are deemed inappropriate or infeasible?
2. Given that the entire rationale for MAIC/STC is premised on the idea that the treatment effects will depend on the population, for which population is the comparison made by MAIC/STC valid, and is this the target population for the decision?

3. Are the uncertainties in the data and assumptions appropriately propagated through to the final estimates?

This TSD also seeks to make recommendations on:

1. The circumstances under which population-adjusted estimates should be used in NICE submissions,
2. The way in which population adjustment should be carried out,
3. Reporting requirements,
4. Priorities for further research

1.4 STRUCTURE OF THIS REPORT

We begin by reviewing the earlier literature on standardisation, generalisation, and calibration (Section 2). Based on this review, we characterise MAIC and STC as "special cases" of the methods developed in the previous literature, pointing out the particular assumptions that are made when MAIC and STC are applied in practice.

In Section 3 we will list the published applications of MAIC and STC, with comments on any specific issues that arise, and with tabulated summaries. In Section 4 we explain our position on the role of population adjustment in the context of submissions to NICE, and outline recommendations on the circumstances in which MAIC and STC might be best applied, along with research recommendations.

The appendices provide additional material, including a flow chart describing the process of making population-adjusted indirect comparisons (Appendix A), proofs and examples of the shared effect modifier assumption (Appendix B), initial suggestions for quantifying residual systematic error in unanchored indirect comparisons (Appendix C), and a worked example of MAIC and STC carried out according to our recommendations complete with accompanying R code (Appendix D).

2. METHODS FOR POPULATION ADJUSTMENT

This section starts with a brief review of earlier literature surrounding population adjustment, including the more recent literature on calibration. We then describe MAIC and STC, noting the similarities and differences with the previous methods, and highlighting the particular assumptions that are made by MAIC and STC. We discuss specific issues which arise from practical application of these methods, focussing on the scenario of submissions to NICE.

2.1 OTHER LITERATURE ON POPULATION ADJUSTMENT

We begin with a brief review of the earlier literature on population adjustment based on propensity score weighting and outcome regression. We then look at a related literature on generalisation of treatment effects, and finally at some recent work on calibration.

2.1.1 PROPENSITY SCORE WEIGHTING

Standardisation is a method closely related to the kinds of adjustment proposed by MAIC and STC. Here, the mean outcomes/responses to be predicted for a target population are based on those observed in an unrepresentative sample, taking into account differences in the distributions of characteristics between the sample and full target population. In our discussion of the methods developed for the standardisation problem, we refer to outcomes under different treatments to remain consistent with our treatment effect calibration scenario; however, in the original context of the standardisation methods, the "treatments" are often exposure classes (e.g. exposed vs. unexposed to a risk factor or intervention) in an observational context, sampled from a larger target population.

Crude direct standardisation, also known as poststratification, subclassification, or direct adjustment, is a basic method of estimating outcomes in a target population of which the sample is an unrepresentative subpopulation, achieved by stratifying the sample population and reweighting the sample means within each subgroup according to the population frequencies.[20] Problems arise with direct standardisation when some subgroups have small (or zero) membership in the sample population, leading to inflated (or even infinite) weights for these subgroups; application is further limited by the number of stratification variables (e.g. age, gender, other clinically relevant risk factors), which must also be categorical (or at least quantised in such a manner).

Rosenbaum[21] proposed a modification known as model-based direct standardisation, in which the weights are found using a parametric model rather than observed population frequencies. Individuals in each subgroup are weighted by the inverse of a *propensity score*, estimated using a logistic model. The propensity score[22] is defined in this context as the conditional probability that an individual from the target population is assigned to the sample given the covariates. When the propensity score model is

correctly specified, weighting removes any imbalance in the distribution of covariates between the sample and target populations. However, incorrect specification of the propensity score model (e.g. wrongly omitting effect modifiers or prognostic variables, higher order terms, or interactions) or the presence of unmeasured effect modifiers or prognostic variables that are in imbalance will result in a biased estimate.

The propensity score has been used in a variety of ways to adjust for imbalances in covariates between a sample population and a larger target population. Propensity score weighting methods in general weight individuals or groups of individuals by the inverse of their propensity score. Differences between the various weighting methods are found in the coarseness of the weights applied: at the finest scale to individuals, at the coarsest scale to whole groups or subclasses, or somewhere in between.[23] MAIC is based on inverse propensity score weighting (IPW), which applies weights at the finest possible scale; each individual in the trial population is given their own weight as a form of Horvitz-Thompson estimator (see Horvitz and Thompson[24]). However, IPW can result in unstable estimates if extreme weights are estimated – a problem not evident in coarser weighting schemes which stabilise the weights. Furthermore, simulation studies have shown that IPW is heavily reliant on correct specification of the propensity score model, and that bias and imprecision are increased by misspecification.[25]

2.1.2 Outcome regression

Outcome regression is an alternative to propensity score weighting. In this method, instead of modelling the propensity score and applying a weighting scheme to the sample individuals, a model for the conditional mean response (or outcome) given treatment and observed covariates is fitted. This model (the *outcome model*) is then used to predict outcomes for each individual in the target population, which are then averaged.

The resulting estimator is unbiased if the outcome model is correctly specified and there are no unmeasured effect modifiers or prognostic variables in imbalance between the populations. Simulation studies have shown that estimators based on a misspecified outcome regression model are less biased and more efficient than estimators based on a misspecified propensity score model,[25] however the associated precisions are overestimated.

2.1.3 Doubly robust estimation

Both propensity score weighting and outcome regression provide methods to estimate outcomes in a target population from a sample subpopulation that differs in covariate balance. However, the estimators are only unbiased if their respective models (for propensity score or outcome) are correctly specified, and if there are no unmeasured effect modifiers or prognostic variables. Doubly robust (DR) estimators

aim to reduce the impact of model misspecification by incorporating both outcome regression and propensity score models into one estimator, which is consistent (and for some estimators unbiased) if at least one of the constituent models is correct.[25-28]

Doubly robust methods "give the analyst two chances" to specify correct models,[28] and demonstrate little loss of efficiency in practice (and sometimes greater efficiency than either method alone). If neither model is correctly specified then the resulting estimator will still be biased and inconsistent.

2.1.4 GENERALISING TREATMENT EFFECTS TO A TARGET POPULATION

There is substantial literature on generalising estimates of relative treatment effects obtained from a RCT into a target population. The methods used are broadly similar to the standardisation literature discussed so far, including propensity score methods[23, 29, 30] and outcome regression. Here, the quantity of interest is the average relative treatment effect in the target population (the population average treatment effect, or PATE) – in contrast to the standardisation literature, which in general is interested in standardising expected outcomes (or absolute effects) to a target population.

The PATE may be estimated by generalising the relative treatment effect from the trial sample – known as the sample average treatment effect (SATE) – into the wider target population. This generalisation is typically effected using similar methods to the standardisation literature discussed above, including propensity score weighting and outcome regression methods. Some authors (e.g. Hartman *et al.*[31]) focus on estimating a related quantity for the treated population only, the population average treatment effect on the treated (PATT), from the sample average treatment effect on the treated (SATT), which is pertinent in some policy decisions. The PATT and SATT are analogous to the PATE and SATE, but the treatment effect is based only on the individuals actually assigned treatment. In a RCT, SATE and SATT are equal due to randomisation (assuming sufficiently large sample size and proper randomisation).

Of particular significance in this literature are the introduction of a rigorous decomposition of the biases in estimating PATE,[32] and tests for generalisability which provide means to verify the assumptions required.

The underlying assumptions required for generalisability and valid estimation of PATE are given by several authors:[23, 31]

1. **Homogeneity of outcomes on each treatment**. Outcomes on treatment and control are the same whether the individual is assigned to the trial or not.
2. **Stable unit treatment value**. The outcomes of one individual are not affected by any other individuals.

3. **Strongly ignorable treatment assignment**. Treatment assignment is random and independent of sample selection from the target population given the observed covariates. This means that there are no prognostic factors or effect modifiers in imbalance between arms of a study.
4. **Strongly ignorable sample assignment**. There are no unmeasured variables related to both sample selection and outcome and, given observed covariates, each individual in the target population has a non-trivial probability (i.e. not zero or one) of being selected into the trial sample.

Assumption 1 may be violated by, for example, protocol differences in inclusion/exclusion criteria.[31] Assumption 2 is met by appropriate study design, and is necessary for causal inference. Assumption 3 is met in RCTs by proper randomisation. Assumption 4 is violated if there are unmeasured effect modifiers or prognostic variables in imbalance between the populations.

In order to assess the assumptions required for generalisability, several authors have proposed what are known as *placebo tests*, suggested by Stuart *et al.*[23] in the context of propensity score models and more generally by Hartman *et al.*[31] In their most general form, placebo tests involve checking whether observed outcomes on a common treatment (not necessarily placebo) in the target population match those predicted by generalisation. The null hypothesis is that there is no difference in the average outcome between populations; however tests of this null hypothesis can have low power, particularly if conditional outcomes by subgroup are investigated, or if the outcome measure has a large variance.[31] An alternative proposition is to use the reverse null hypothesis that there is a difference in average control outcome between populations, a test of which will then only support generalisability if there is sufficient evidence and sufficient power to reject the null.[33] Tests of this form involve specifying a cut-off value; differences smaller than this mean that the control outcomes between populations are considered equivalent.

Placebo tests can demonstrate failures of assumptions 1, 2, and 4 above, however they cannot ascertain which assumption or assumptions are violated, nor can they detect multiple violations whose resulting biases cancel each other out.[31] Furthermore, a placebo test comparing observed and predicted placebo outcomes only has capacity to check for unobserved prognostic variables in imbalance in assumption 4 and *not* for unobserved effect modifiers in imbalance. A placebo test comparing observed and predicted outcomes on a common active comparator (if available) would additionally be able to detect unobserved effect modifiers in imbalance (but would not be able to discern whether the unobserved covariate was an effect modifier or prognostic variable).

When propensity score methods are used to generalize relative treatment effects, Stuart *et al.*[23] suggest examining the difference in average propensity scores between the trial population and target

population; a difference in the mean propensity score greater than 0.25 standard deviations indicates that the generalisation is largely based on extrapolation, and will be heavily dependent on the propensity score model used.

2.1.5 CALIBRATION OF TREATMENT EFFECTS

The literature reviewed thus far seeks to generalise either the absolute outcomes (Sections 2.1.1-2.1.3) or the relative treatment effects (Section 2.1.4) observed in a sample sub-population, under some strict assumptions, to those that would be observed in a target population. There has however been no attempt to perform indirect treatment comparisons in the target population, which is our problem of interest. Additionally, we now wish to consider the sample and target populations as distinct and independent (e.g. from two non-overlapping clinical trials), whereas previously the sample was considered an unrepresentative subpopulation of the target population. Several authors have framed this as a treatment effect calibration problem, where information on treatment effects and covariates in one population is used to estimate treatment effects in another population with different known covariate values,[34-37] and note that it is similar to the generalisation problem (Section 2.1.4).

The work on calibration assumes that IPD are available on both the AB and AC trials, so the methods proposed are not strictly relevant to the problem that MAIC and STC set out to address. However, we review this literature here because it contains some clear statements of the assumptions made by MAIC and STC.

Recently there has been a specific interest from the US Food and Drug Administration (FDA) in calibration methods for the analysis of non-inferiority studies, which compare a treatment B with an active comparator A and thus lack a placebo arm C. When the quantity of interest is treatment effect relative to placebo, a historical placebo-controlled trial with the active comparator (i.e. an AC trial) may be used to calibrate the treatment effect by estimating the placebo effect that would be observed in the AB trial (known as a *putative placebo analysis*). Calibration methods are interested in estimating the average relative treatment effect of B vs. C in the AB population on an appropriate scale, which can be done in one of several ways depending on the assumptions one is willing to make. (This is in contrast with MAIC/STC, where the target of inference is the B vs. C effect that would be observed in the AC trial if the B arm was included.)

The first possibility is an approach based on the assumption of *constancy of absolute effects*, which requires that there are no prognostic variables or effect modifiers in imbalance between the two populations. This is of course absurd, as there is no randomisation between trials – only within. No

accepted methods for evidence synthesis or indirect comparison, whether population-adjusted or not, make this impossibly strong assumption (see Table 2).

Another possibility is an approach based on the assumption of *conditional constancy of absolute effects* (also known as *treatment-specific conditional constancy* in the calibration literature). This means that the expected absolute outcomes under treatment C are identical between the two trial populations at any given set of covariate values. This assumption is very strong (if not implausibly so), as it requires all effect modifiers and prognostic variables to be available.[37] Estimation of the indirect comparison under this assumption proceeds via one of the previously discussed methods (e.g. propensity score weighting, outcome regression), which is used to predict absolute outcomes in the AB population. We note that conditional constancy of absolute effects is equivalent to ignorable sample assignment as described in the generalisation literature (assumption 4, Section 2.1.4).

To avoid making such a strong assumption about prognostic variables, inferences could be made instead using an assumption of *constancy of relative effects* (sometimes referred to simply as *constancy*), meaning that the relative C vs. A effect observed in the AC trial is identical to that which would be observed in the AB trial. However this is often questionable, as constancy of relative effects requires that all effect modifiers (whether measured or unmeasured) are perfectly balanced between the two trial populations. This is akin to the consistency assumption (on the transformed scale) that is standard in NMA:[38] consistency is assumed to hold exactly for a fixed effect analysis, and is relaxed in a random effects analysis where consistency is only assumed to hold in expectation. The consistency assumption in random effects models is reasonable when contrasts are informed by many trials, allowing the impact of effect modifiers to "balance out", but less so in sparse networks. Development of population adjustment for the very sparse networks of comparisons often seen in submissions to NICE is therefore well motivated.

Instead of making any of the three strong assumptions above, calibration methods rely on an assumption of *conditional constancy of relative effects* (sometimes referred to simply as *conditional constancy*). This states that the relative C vs. A effect observed in the AC trial at a given covariate value (e.g. the effect at age 55) is equal to the C vs. A effect which would be observed in the AB trial at that same covariate value. This assumption may be more valid, as only effect modifiers are required to be adjusted for; estimators based on the conditional constancy of relative effects assumption respect randomisation which balances prognostic variables within studies.

Calibration methods have been proposed in various forms: covariate adjustment, which is a form of outcome regression;[34] likelihood reweighting, which is a form of propensity score weighting;[36] and

doubly robust methods.[37] Another estimator recently proposed by Zhang *et al.*[37] is known as a *conditional effect* (CE) estimator, which models the conditional relative treatment effect directly rather than modelling (transformed) outcomes, and may also be combined into doubly robust estimators. In practice, all of the above methods require IPD on the historical AC trial in order to infer comparisons in the AB population; this differs from the calibration scenarios into which MAIC and STC have been proposed, where IPD on the AC trial are unavailable and comparisons are inferred in the AC population. Zhang[34] notes that covariate adjustment may be performed using aggregate data from the AC trial if the coefficients in the outcome regression and their covariance matrix are published, although this seems unlikely.

2.2 POPULATION ADJUSTMENT WITH LIMITED IPD

The core principles of MAIC and STC remain the same as in the general calibration literature, however the problem scenario is modified slightly: rather than individual patient data (IPD) being available in all study populations, IPD are only available in the AB trial, with aggregate data in the AC trial along with information on the covariate distribution. Ideally the full joint distribution of $\boldsymbol{X}$ is known, but frequently in practice only the marginal mean and standard deviation of each covariate is known. Due to the lack of IPD from the AC trial, standard approaches to fitting both propensity score and outcome models may not be used. We outline both MAIC and STC approaches below. A worked example of MAIC and STC as conforming to our recommendations is included in Appendix D.

2.2.1 MATCHING-ADJUSTED INDIRECT COMPARISON (MAIC)

MAIC is a form of the non-parametric likelihood reweighting method previously discussed in our review of the calibration literature (Section 2.1.5), which allows the propensity score logistic regression model to be estimated without IPD in the AC population. The mean outcomes $\mu_{t(AC)}$ on treatment $t = A, B$ in the AC target population are estimated by taking a weighted average of the outcomes $Y_{it(AB)}$ of the $N_{t(AB)}$ individuals in arm t of the AB population

$$\hat{Y}_{t(AC)} = \frac{\sum_{i=1}^{N_{t(AB)}} Y_{it(AB)} w_{it}}{\sum_{i=1}^{N_{t(AB)}} w_{it}}, \tag{6}$$

where the weight w_{it} assigned to the i-th individual receiving treatment t is equal to the odds of being enrolled in the AC trial vs. the AB trial. Conceptually this is very similar to the previously discussed inverse propensity weighting method in the standardisation literature (Section 2.1.1). As with likelihood reweighting (from which MAIC is derived), the weights themselves are estimated using logistic

regression as $\log(w_{it}) = \alpha_0 + \boldsymbol{\alpha}_1^T \boldsymbol{X}_{it}$, where $\boldsymbol{X}_{it}$ is the covariate vector for the i-th individual receiving treatment t; however, the regression parameters are not estimable using standard methods due to the lack of IPD in the AC trial, in particular a lack of information on the joint distribution of covariates. If the joint covariate distribution was available in the AC trial, then the likelihood reweighting approach of Nie *et al.*[36] would be feasible, with the possibility of the sufficient statistics replacing the full IPD. Because only marginal information is available, Signorovitch *et al.*[3] propose using a method of moments to estimate $\hat{\boldsymbol{\alpha}}_1$ so that the weights exactly balance the mean covariate values (and any included higher order terms, for example squared covariate values to balance the variance) between the weighted AB population and the AC population. When $\bar{\boldsymbol{X}}_{(AC)} = \boldsymbol{0}$, Signorovitch *et al.* show that this is equivalent to minimising $\sum_{t=A,B} \sum_{i=1}^{N_{t(AB)}} \exp\left(\boldsymbol{\alpha}_1^T \boldsymbol{X}_{it}\right)$. The estimator in equation (6) is then equal to

$$\hat{Y}_{t(AC)} = \frac{\sum_{i=1}^{N_{t(AB)}} Y_{it(AB)} \exp\left(\hat{\boldsymbol{\alpha}}_1^T \boldsymbol{X}_{it}\right)}{\sum_{i=1}^{N_{t(AB)}} \exp\left(\hat{\boldsymbol{\alpha}}_1^T \boldsymbol{X}_{it}\right)},$$

noting that $\exp(\hat{\alpha}_0)$ cancels from the top and bottom of the fraction. Anchored and unanchored indirect comparisons are then formed using equations (4) and (5) respectively. Although MAIC can be used to facilitate indirect comparisons on any scale, the MAIC literature almost exclusively performs comparisons on the natural outcome scale (i.e. with $g(\cdot)$ the identity function). Typically, standard errors for MAIC estimates are calculated using a robust sandwich estimator[39] (see the appendix of Signorovitch *et al.*[3]). Sandwich estimators are derived empirically from the data rather than making overly strong assumptions about the weights, to account for the fact that the weights are estimated rather than fixed and known. Signorovitch *et al.*[3] suggest that the effective sample size (ESS) of the pseudo-population formed by weighting the AB population is approximated by

$$\text{ESS} = \left(\sum_{t=A,B} \sum_{i=1}^{N_{t(AB)}} \hat{w}_{it}\right)^2 \Big/ \sum_{t=A,B} \sum_{i=1}^{N_{t(AB)}} \hat{w}_{it}^2 \,. \qquad (7)$$

This approximate ESS is only accurate if the weights are fixed and known, or if they are uncorrelated with outcome – neither of which is true here; as such, this approximation is likely to be an underestimate of the true ESS.[40] However, small effective sample sizes are an indication that the weights are highly variable due to a lack of population overlap, and that the estimate may be unstable. The distribution of weights themselves should also be examined directly, to diagnose population overlap and to highlight any overly influential individuals. It is not possible to apply traditional propensity score tools for "balance checking" here, as propensity scores are only estimated for the AB trial, and the method of

moments by definition ensures covariate balance (at least in the means, and up to the level of information published in the AC trial).

2.2.2 Simulated Treatment Comparison (STC)

STC is a modification of the covariate adjustment method previously discussed in our review of the calibration literature (Section 2.1.5). Firstly, an outcome model is fitted using the IPD in the AB trial:

$$g\left(\mu_{t(AB)}\left(\boldsymbol{X}\right)\right)=\beta_0+\boldsymbol{\beta}_1^T\boldsymbol{X}+\left(\beta_B+\boldsymbol{\beta}_2^T\boldsymbol{X}^{EM}\right)I\left(t=B\right) \quad (8)$$

where β_0 is an intercept term, $\boldsymbol{\beta}_1$ is a vector of coefficients for prognostic variables, β_B is the relative effect of treatment B compared to A at $\boldsymbol{X}=\boldsymbol{0}$, $\boldsymbol{\beta}_2$ is a vector of coefficients for effect modifiers $\boldsymbol{X}^{EM}$ (a subvector of the full covariate vector $\boldsymbol{X}$), and $\mu_{t(AB)}\left(\boldsymbol{X}\right)$ is the expected outcome of an individual assigned treatment t with covariate values $\boldsymbol{X}$ which is transformed onto a chosen linear predictor scale with link function $g\left(\cdot\right)$.

The model in equation (8) is a more general form of that given by Ishak *et al.*[4], which does not include any effect modifier terms. Ishak *et al.* then form (on the natural outcome scale) either an unanchored indirect comparison $\hat{\Delta}_{BC(AC)}=\bar{Y}_{C(AC)}-\hat{Y}_{B(AC)}$, or an anchored indirect comparison $\hat{\Delta}_{BC(AC)}=\bar{Y}_{C(AC)}-\bar{Y}_{A(AC)}-\left(\hat{Y}_{B(AC)}-\hat{Y}_{A(AC)}\right)$, where $\hat{Y}_{A(AC)}$ and $\hat{Y}_{B(AC)}$ are predicted from the outcome regression by substituting in mean covariate values to obtain $\hat{Y}_{A(AC)}=g^{-1}\left(\hat{\beta}_0+\hat{\boldsymbol{\beta}}_1^T\bar{\boldsymbol{X}}_{(AC)}\right)$ and $\hat{Y}_{B(AC)}=g^{-1}\left(\hat{\beta}_0+\hat{\boldsymbol{\beta}}_1^T\bar{\boldsymbol{X}}_{(AC)}+\hat{\boldsymbol{\beta}}_B+\hat{\boldsymbol{\beta}}_2^T\bar{\boldsymbol{X}}_{(AC)}^{EM}\right)$. Ishak *et al.* note that these estimators (and hence an indirect comparison on the natural outcome scale based on them) are systematically biased whenever $g\left(\cdot\right)$ is not the identity function (i.e. not $g\left(y\right)=y$), because the mean outcome depends on the full distribution of the covariates and not just their mean. Instead of substituting in mean covariate values in this case, Ishak *et al.* suggest that estimates are obtained by first drawing samples from the joint covariate distribution in the AC trial and then averaging over the predicted outcomes based on the regression model. (We discuss how this is typically achieved using published data, and the additional assumptions required, in Section 2.3.4.) This simulation approach however introduces additional variation as, rather than computing an average over the distribution of covariates in the AC population, the estimated quantity is now the expected effect for a randomly selected individual from the AC population (i.e. the predictive distribution), leading to an underestimate of the precision of the final indirect comparison estimate.

Forming indirect comparisons directly on the natural outcome scale, as advocated by the STC literature and described above, causes several problems (see Section 2.3.1.2). To avoid these, we strongly recommend that anchored and unanchored indirect comparisons are formed on the linear predictor scale using equation (4) or (5) respectively. Standard tools for model checking (such as AIC/DIC, examining residuals, etc.) may be used when constructing the outcome model in the AB trial; however (as with MAIC), additional assumptions are required to predict outcomes in the AC population, which are difficult to test when there is little data available.

Whilst the above formulation of STC is seen in Ishak *et al.*[4] and in all the published applications of STC to date, an earlier paper[6] suggests that an indirect comparison may be performed in the AB population via extension to the above steps. We have not identified any applications employing this method.

2.2.3 NETWORK META-REGRESSION WITH LIMITED IPD

If individual patient data are available on both the AB and AC studies, a network meta-regression using IPD is the gold standard approach.[12, 41-44] There has, understandably, been interest in generalising network meta-regression to situations where only limited IPD are available in a network of treatment comparisons; our scenario with one AB IPD study and one AC aggregate study is then a special case. Currently, there are two main forms of network meta-regression which combine both IPD and aggregate data, which primarily differ in how the regression model is defined at the individual level and at the aggregate level. We discuss both approaches here, in the context of our two-study scenario.

The first approach builds upon that of Sutton *et al.*[45] for pairwise meta-regression.[7-10] Two regression models are fitted simultaneously, one describing individual level outcomes in the AB trial, and another describing the aggregate outcome in the AC trial:

Individual:

$$g\left(\mu_{t(AB)}\left(\boldsymbol{X}\right)\right) = \beta_{0(AB)} + \boldsymbol{\beta}_1^T \boldsymbol{X} + \left(\beta_B + \boldsymbol{\beta}_2^T \boldsymbol{X}^{EM}\right) I\left(t = B\right)$$

Aggregate:

$$g\left(\mu_{t(AC)}\right) = \beta_{0(AC)} + \left(\beta_C + \boldsymbol{\beta}_2^T \bar{\boldsymbol{X}}_{t(AC)}^{EM}\right) I\left(t = C\right) \quad (9)$$

Due to the lack of data, there are some restrictions on the more general models which have been proposed for larger networks:[8, 9] a fixed effects model must be used, and the treatment by effect modifier interaction coefficient $\boldsymbol{\beta}_2$ must be shared between treatments B and C and between the individual and aggregate level. This second restriction is at first glance akin to the shared effect modifier assumption discussed later in Section 4.1.3, although on further inspection it is far stronger – the effect modifier is required to act in the same manner on both the aggregate level and on the individual level.

This assumption is only valid if the identity link is used and all effect modifiers are accounted for (and proper randomisation has occurred); imposing this assumption when it does not hold results in aggregation bias (a form of ecological bias).[8, 9, 41, 46]

The second approach derives from a type of model proposed by Jackson *et al.*[47, 48] known as *hierarchical related regression*. This model avoids the pitfalls of the first by correctly relating the individual and aggregate levels so that aggregation bias does not occur. The basic idea is a natural one; the aggregate data arise from averaging over a population of individuals, so the aggregate level model arises from averaging (i.e. integrating) the individual model over a population. The resulting model may be written in most general form as

Individual:

$$g\left(\mu_{t(AB)}\left(\boldsymbol{X}\right)\right) = \beta_{0(AB)} + \boldsymbol{\beta}_1^T \boldsymbol{X} + \left(\beta_B + \boldsymbol{\beta}_2^T \boldsymbol{X}^{EM}\right) I\left(t = B\right)$$

Aggregate: (10)

$$\mu_{t(AC)} = \int_{\boldsymbol{X}_{(AC)}} g^{-1}\left(\beta_{0(AC)} + \boldsymbol{\beta}_1^T \boldsymbol{X} + \left(\beta_C + \boldsymbol{\beta}_2^T \boldsymbol{X}^{EM}\right) I\left(t = C\right)\right) f_{(AC)}\left(\boldsymbol{X}\right) d\boldsymbol{X}$$

where $f_{(AC)}\left(\boldsymbol{X}\right)$ is the joint distribution of $\boldsymbol{X}$ in the AC trial population. If the full joint distribution is not available for the AC trial (as is likely with published data), an approximation may be used – for example by assuming a normal distribution (or another appropriate distribution, such as log normal) for continuous covariates with the reported mean and standard deviation, and either imputing correlations between covariates from the AB trial or assuming that they are zero. Note that this model reduces to the gold-standard IPD network meta-regression when IPD are available for all studies, and is equally applicable for analysing larger networks of treatments with a mixture of IPD and aggregate data available. When used in our simple two-study scenario, model (10) does require the shared effect modifier assumption in order to estimate the parameters due to lack of data; however, this assumption may not be required when a larger network of studies is available, or perhaps if external information on the effect modifiers of treatment C is available. Model (10) is equivalent to model (9) if an identity link is used and all effect modifiers are accounted for.

The individual level model here is of the same form as above in model (9). The aggregate level model however is found by integration of this individual level model, and therefore may not be straightforward to explicitly write down. Jansen[7] describes a special case of model (10) for the simple case of a binary outcome and binary covariates. When all covariates are binary (or categorical), it is simple to rewrite the integration as a sum over each level of the covariates, so that the aggregate level model becomes

$$\mu_{t(AC)} = \sum_{\boldsymbol{X}_j} g^{-1}\left(\beta_{0(AC)} + \boldsymbol{\beta}_1^T \boldsymbol{X}_j + \left(\beta_C + \boldsymbol{\beta}_2^T \boldsymbol{X}_j^{EM}\right) I\left(t = C\right)\right) f_{(AC)}\left(\boldsymbol{X}_j\right) \qquad (11)$$

where $\boldsymbol{X}_j$ is a discrete level of the covariates, and $f_{(AC)}\left(\boldsymbol{X}_j\right)$ is simply the proportion of AC trial individuals in the category $\boldsymbol{X}_j$. We are not currently aware of any more general applications of model (10) in the literature; in the absence of a more sophisticated approach, model (11) may be used to incorporate continuous covariates by splitting them into discrete categories (e.g. splitting ages into 5 year bands), at the expense of loss of information.

The hierarchical network meta-regression approach in model (10) represents an alternative class of methods to those such as MAIC and STC. The hierarchical approach models individual-level relationships and is able to provide internally consistent inferences at both the individual level and at an aggregate level like a standard indirect comparison. Methods such as MAIC and STC use IPD to predict average outcomes on study arms, and then effect the indirect comparison at the aggregate study level. We could therefore refer to MAIC and STC as forms of *population-adjusted study-level indirect comparisons*, and the hierarchical approach as a form of *population-adjusted individual-level indirect comparison*. Despite the apparent benefits of the hierarchical approach, we focus on MAIC and STC for the remainder of this report. We do however expect many of the properties of STC to hold for these methods, and the recommendations made in Section 4.2 are applicable to general forms of population adjustment including those based on network meta-regression as well as MAIC and STC. We comment further on network meta-regression for mixed IPD and aggregate data in Section 4.3.

2.2.4 Other forms of population reweighting

The application of weights to individuals in the IPD population in order to balance the covariate distributions between trials is a general technique which we shall refer to as *population reweighting*. MAIC as described in Section 2.2.1 is currently the most widely used form of population reweighting when IPD are only available for the AB trial. Another form of population reweighting is based on entropy balancing,[49] and was first suggested for treatment effect calibration by Belger *et al.*[50, 51] Rather than seeking to estimate a propensity score with which to create weights, entropy balancing methods are designed to estimate weights by directly matching moments of the covariate distributions (such as the mean and standard deviation). As MAIC uses the method of moments to estimate weights, the methods up to this point are effectively identical. However, entropy balancing methods apply an additional constraint when estimating the weights; the optimal entropy balancing weights are those which are as close as possible to uniform weights (that is, as close as possible to no weighting at all). This additional constraint means that entropy balancing methods should have equal or reduced standard error compared to MAIC, whilst achieving the same reduction in bias.

Different schemes for applying weights have also been proposed. MAIC, as described in Section 2.2.1, estimates weights for the entire AB population at once to balance covariate distributions with the entire

AC population. Belger *et al.*[50, 51] compare anchored and unanchored MAIC with other possible approaches, which involve splitting apart trial arms and balancing covariate distributions separately between the control arms (A) and between the treatment arms (B and C) in the IPD and aggregate populations. The properties of such "splitting" approaches in comparison with a more typical population reweighting are largely unknown, and require further investigation. For this reason we do not comment further on these approaches in this TSD.

2.3 ASSUMPTIONS AND PROPERTIES OF MAIC AND STC IN ANCHORED AND UNANCHORED COMPARISONS

We now examine in detail the assumptions made by MAIC and STC which are required to achieve a valid indirect comparison in the target population. If these assumptions are violated, the resulting estimate may be biased. It is critical to observe that the necessary assumptions differ between the anchored and unanchored forms of indirect comparison (equations (4) and (5) respectively), with the unanchored indirect comparison requiring stronger assumptions. We do not discuss the first three core assumptions specified in the generalisation literature (homogeneity of effects, stable unit treatment value, and ignorable treatment assignment), as they must generally be assumed to hold for any form of indirect comparison or meta-analysis. If there are issues with the randomisation within studies (violating assumption 3 in Section 2.1.4) then these may be addressed prior to MAIC/STC analysis by the application of typical weighting/regression adjustment methods.

2.3.1 ANCHORED COMPARISONS

The MAIC and STC literature typically advocates performing indirect comparisons directly on the outcome scale, with $g(\cdot)$ the identity function in equation (4) for an anchored comparison, so that

$$\hat{\Delta}_{BC(AC)} = \bar{Y}_{C(AC)} - \bar{Y}_{A(AC)} - \left(\hat{Y}_{B(AC)} - \hat{Y}_{A(AC)}\right). \tag{12}$$

2.3.1.1 MAIC, and STC with a linear model

When making an anchored indirect comparison in the AC population on the outcome scale as in equation (12), both MAIC and STC (using a linear outcome model with identity link) rely on an assumption of conditional constancy of relative effects on the outcome scale – that the differences in the relative effects that would be observed between studies are entirely accounted for by an imbalance in the effect modifier variables $\boldsymbol{X}^{EM}$ (see Section 2.1.5). The implication of this assumption is that $\boldsymbol{X}^{EM}$ must contain every effect modifier that is in imbalance between the two studies, otherwise the indirect comparison is still biased. Note that both effect modifiers and conditional constancy of relative effects here are defined on the outcome scale due to the indirect comparison being made on this scale.

STC requires the correct specification of the form of the outcome model in order to provide unbiased estimates. When an anchored comparison is made, an unbiased estimate is still obtained even if some or all prognostic variables (that aren't also effect modifiers) are omitted from or misspecified in the model (and an intercept term is included). However, inclusion of prognostic variables in the outcome model should in theory lead to more precise estimation of the treatment effect and effect modifier parameters within the model and the resulting indirect comparison, as a portion of the variability is accounted for by the prognostic variables.

In the present MAIC literature,[3-5] there is no discussion of which variables (prognostic and/or effect modifying) should be included in the weighting model; the prevailing choice in applications of MAIC to date appears to be to include as many variables as possible, regardless of effect modifier status or level of imbalance (see Section 3). However, the choice of variables to be matched/weighted on should be carefully considered: including too many variables will reduce the effective sample size, negatively affecting the precision of the estimate; conversely, failure to include relevant variables will result in a biased estimate. Therefore, for an anchored indirect comparison, the weighting model must include all effect modifiers (both those in balance and imbalance between the studies), but no prognostic variables. Including effect modifiers that are already balanced in the weighting model ensures that they remain imbalanced after the weighting, and there will be negligible impact on the standard error due to their inclusion. Imbalances in prognostic variables are taken care of by the randomisation within studies (and the subsequent "adjustment" to the comparison with the control arms), and their inclusion in the matching model only reduces the effective sample size.

2.3.1.2 STC with a non-identity link

In the case that STC is carried out with a non-identity link function, there arises a conflict of scale when equation (12) is used to form an indirect comparison on the natural outcome scale: the outcome model defines a specific transformed linear predictor scale, upon which additivity is assumed and effect modifiers and prognostic variables are defined, whereas the indirect comparison is formed on the natural outcome scale. Effect modifier status is mathematically demonstrable to be scale-specific (e.g. Brumback and Berg[52]), and the status of a variable as an effect modifier on one scale does not imply (either positively or negatively) the effect modifier status on any other scale. Therefore, performing the indirect comparison on one scale whilst fitting the outcome model on another raises questions about the interpretation of the model and of the indirect comparison.

The advantage of an anchored indirect comparison over an unanchored indirect comparison is also in doubt in this case, as the aim of cancelling out prognostic variables on the outcome scale in the anchored indirect comparison is in contradiction with their definition on the linear predictor scale in the outcome model. It is unclear at present whether the anchored comparison leads to a reduction in bias and reliance

on model specification or an increase, compared to the unanchored comparison. However, it is clear that, as prognostic variables (defined on the linear predictor scale) will not cancel in the anchored indirect comparison (defined on the outcome scale), any misspecification or omission of prognostic variables in the outcome model will lead to a biased estimate. Therefore, an indirect comparison made using STC with a non-identity link makes the assumption that $\boldsymbol{X}$ contains both all effect modifiers and all prognostic variables (i.e. conditional constancy of absolute effects) with respect to the linear predictor scale, and that the outcome model is correctly specified.

Performing the indirect comparison on the transformed linear predictor scale as in equation (4) (instead of the outcome scale) would eliminate these concerns, and once again lead to reliance upon the weaker assumption of conditional constancy of relative effects. This is the usual method employed in standard indirect comparisons.[1, 2] We discuss the choice of scale further in Section 2.3.3.

2.3.2 UNANCHORED COMPARISONS

If an unanchored comparison is made (equation (5)), whether on the outcome scale or transformed scale, then both MAIC and STC rely on the conditional constancy of absolute effects assumption; the differences between absolute outcomes that would be observed in each trial are entirely explained by imbalances in prognostic variables and effect modifiers $\boldsymbol{X}$ with respect to the chosen scale. Under this assumption, $\boldsymbol{X}$ must contain both every prognostic variable and every effect modifier that is in imbalance between the two studies – an assumption that is largely deemed unreasonable (if it were, there would be no reason to undertake randomised controlled trials). Conditional constancy of absolute effects may be partially assessed in a connected scenario through the use of placebo tests using the common comparator (see Section 2.1.4). If the conditional constancy of absolute effects assumption fails then the unanchored estimator is invalid and an anchored estimator making use of the conditional constancy of relative effects assumption should be used. However, such tests cannot be used to justify an unanchored comparison for two reasons: (i) lack of statistical power; and (ii) conditional constancy of absolute effects is only partially assessed if the common comparator is placebo, as residual imbalances in observed or unobserved effect modifiers cannot be evaluated. It should also be noted that, whilst the traditional approach is to adjust for all available variables, these may nevertheless be limited (especially in the published aggregate data), and therefore such an approach alone is not sufficient justification for the conditional constancy of absolute effects assumption.

STC furthermore assumes that the outcome model is correctly specified in both prognostic variables and effect modifiers; it is thus more burdensome to specify an outcome model for an unanchored comparison than for an anchored comparison, as the prognostic variables and their model specification become critical in the unanchored case. The impact of performing an unanchored indirect comparison

on a different scale to that of the linear predictor is currently unknown, although the concerns over interpretability raised for the anchored case in Section 2.3.1.2 still stand.

If a MAIC is to be performed, the weighting model must include every effect modifier and prognostic variable – compared to the anchored case, where only effect modifiers are required. An immediate consequence of this is that an unanchored indirect comparison performed using MAIC will always have less precision than an anchored indirect comparison using MAIC in the presence of an imbalance of prognostic variables, and – more importantly – is more likely to be biased given that all prognostic variables in imbalance must be included in the weighting model as well as effect modifiers.

2.3.3 CHOICE OF SCALE FOR INDIRECT COMPARISON

The standard practice for indirect comparisons is that they are made on a transformed scale (e.g. on the log scale for odds ratios and risk ratios), rather than on the natural outcome scale;[1, 2] for the purposes of a CEA, the resulting estimates may be back-transformed onto the (possibly more interpretable) natural scale. The reasons for this choice include approximate normality and the stabilisation of variance, however the most critical reason with regards to indirect comparisons is that effects are assumed to be additive and linear on the transformed scale. Therefore the apparently pervasive choice amongst present applications of MAIC and STC to perform comparisons directly on the natural outcome scale in the face of a more usual transformed scale is disconcerting, and somewhat of a contradiction of assumptions. We cannot be certain of the impact of such conflicts of scale without comprehensive simulation studies.

This is made most clear by STC when an outcome model is (quite correctly) specified with a non-identity link function (see Section 2.3.1.2): the outcome model defines effects linearly and additively on the transformed linear predictor scale, which is in direct contradiction with the subsequent assumption of linearity and additivity on the outcome scale used by the indirect comparison. Furthermore, the definition and interpretation of effect modifiers and prognostic variables is entirely scale-specific, and results in conflicts and contradictions when the outcome model and indirect comparison are on differing scales.

A potential and oft-cited advantage of MAIC is that it is perceived to be "scale-free", in the sense that the definition of the weighting model does not require any fixed outcome scale to be chosen.[3, 4] We however express caution at this notion: it is true that no outcome model need be assumed to create the weighting model, but the subsequent indirect comparison does assume additivity on a specific scale, and therefore neither MAIC nor STC are "scale-free" in this important sense.

2.3.4 IMPACT OF HAVING ACCESS TO ONLY MARGINAL COVARIATE DISTRIBUTION

Thus far we have considered MAIC and STC in the scenario where, despite not having access to IPD on the AC trial, sufficient information on the joint covariate distribution is available. In practice even this level of detail is unlikely, as published trials frequently report only details of the marginal covariate distributions (e.g. mean/median and standard deviation for continuous covariates, or proportion of individuals with a binary/categorical trait). This leads to an additional assumption being required for both MAIC and STC: either that (i) the joint distribution of covariates in the AC trial is the product of the (published) marginal distributions, or (ii) the correlations between covariates in the AC trial are the same as those observed in the AB trial.

This assumption is most explicit when STC is used. Ishak *et al.*[4] propose that, in order to create predictions into the AC population, missing correlations between covariates in the AC population are assumed to be the same as those observed in the AB population.

MAIC does not explicitly specify any form of outcome model, however there is an implicit outcome model which is inferred when the indirect comparison is formed. Specifically, effects are assumed to be additive on the scale of the indirect comparison, as are the actions of effect modifiers and prognostic variables. When covariate correlations are not available from the AC population (and therefore cannot be balanced by inclusion in the weighting model), they are assumed to be equal to the correlations amongst covariates in the pseudo-population formed by weighting the AB population.

However, if an anchored indirect comparison is made (from either MAIC or STC), then, due to the cancellation of prognostic variables, only correlations amongst effect modifiers will affect the indirect comparison, and the assumption of identical correlations amongst prognostic variables between the two trial populations can be dropped. Furthermore, if there are no multi-way treatment by effect modifier interactions in the (for MAIC, implicit) outcome model (or any interactions at all, for an unanchored comparison), then the estimated indirect comparison will remain unbiased even if the correlations between covariates differ between the two trial populations.

2.3.5 CHOICE OF TARGET POPULATION

The premise of both MAIC and STC is that the treatment effect depends on the population. It is therefore not sufficient to use MAIC or STC to generate an "unbiased" comparison in just any population; they only achieve this purpose if they can produce a fair comparison in the target population for the decision. In general, the target population should be a UK cohort or registry study population relevant to the clinical decision, which is unlikely to match the population of the AC trial. However, MAIC and STC as currently proposed are unable to achieve estimates in any population other than that of the AC study.

We present an extension in Section 4.2.7 which enables indirect comparisons to be made in any target population, given an additional assumption.

The population-specific nature of MAIC and STC analyses can lead to apparently contradictory conclusions being drawn from the same pair of trials, simply by taking the alternate company's perspective and swapping the roles of the AB and AC studies, having instead IPD on the AC trial and aggregate data on the AB trial. This problem has already arisen in analyses from competing companies: Novartis and AbbVie presented MAIC analyses of the same two trials comparing secukinumab and adalimumab to placebo as treatments for ankylosing spondylitis (AS).[53, 54] Each company had IPD on their own trial, but not on their competitor's trial. The results from each company's MAIC appear to be in conflict, with one company claiming significant differences in efficacy in favour of secukinumab, and the other claiming comparable efficacy but improvements in cost effectiveness for adalimumab. Importantly we note that, as MAIC (and STC) attempts to produce estimates in the AC population, the two MAIC analyses are aiming to provide estimates in two different target populations – the population of the competitor's trial in each case. Furthermore, the Novartis trial population included both treatment experienced and treatment naïve patients, whereas the AbbVie trial population included only treatment naïve patients. Due to the lack of population overlap concerning treatment experienced patients, it is impossible for a MAIC from AbbVie's perspective to generate estimates for the full Novartis trial population. However, even if both trial populations overlapped perfectly, we would still expect there to be differing estimates depending on which company's perspective is taken – precisely because the two study populations have been deemed incomparable directly due to an imbalance in effect modifiers; if there were no such imbalance, then there would be no need to conduct an anchored indirect comparison instead of the usual indirect comparison. The real conflict, therefore, lies not in the results produced by the two MAICs, but in deciding which of the two study populations better represents the true target population. Ironically, each company is left in the position of implicitly assuming that their competitor's trial is more representative than their own.

This prospect of conflicting estimates from different companies becomes exponentially worse as MAIC/STC is extended to multiple trials and multiple treatments. For example in a star-like structure of AB, AC, AD, AE studies, if each company performed a MAIC/STC using IPD available on their own trial, and effect modification was present, they would generate among them four incoherent sets of three pair-wise indirect comparisons, none of which could be compared to each other.

2.3.6 SAMPLING VARIATION IN THE TARGET POPULATION

MAIC and STC, as currently portrayed, produce estimates of mean outcomes on each treatment in the AC study sample, rather than in the AC population. In other words, the sampling uncertainty of the AC trial sample is ignored.

There is substantial literature on super-population average treatment effects (SPATE), which addresses precisely this issue (for an introduction, see Imbens and Rubin,[55] chapter 6). In the context of our calibration scenario, the AB and AC trials are seen as samples from a larger super-population (the true target population), and the estimates in the AC trial can be turned into estimates in the target population by accounting for the additional sampling variation. A notable special case occurs when the inclusion/exclusion criteria of the AC trial match exactly the true target population and the individuals enrolled in the AC trial are randomly sampled from the true target population; then the point estimates provided by MAIC or STC in the sample population are exactly carried over to the true target population, with an increase in standard error reflecting the sampling uncertainty.

2.4 UNCERTAINTY PROPAGATION

We break down the uncertainty in the estimates resulting from MAIC and STC into three sources: sampling variation within the studies, uncertainty due to the imbalance in covariate distributions, and uncertainty due to estimation of the weighting/outcome model. Both MAIC and STC fully account for the sampling variation within the studies, and propagate this through to the final estimate.

MAIC inherently accounts for the uncertainty due to the imbalance in covariate distributions: greater differences between the covariate distributions lead to an increase in the variation of weights (some become larger, some become smaller) and hence a reduction in effective sample size. Standard errors for MAIC estimates are typically obtained using robust sandwich estimators,[3] which account for the fact that the weights are estimated rather than fixed and known. Alternative methods for incorporating all sources of uncertainty in MAIC include bootstrapping techniques,[56] or incorporating the analysis in a Bayesian framework.

Whether or not STC takes into account the latter two sources of variation depends upon how the predicted outcomes into the AC study are treated. If the predicted outcomes are treated as fixed and known (as if they had actually been observed), then the estimates resulting from STC will not take into account either the uncertainty due to covariate imbalance (which may lead to extrapolation if there is insufficient overlap between the two populations), or due to the estimation of the outcome model parameters. However, if the predicted outcomes are correctly considered along with their associated prediction error, then the resulting estimates will account for all three sources of variation.

2.5 CALIBRATING POPULATION-ADJUSTED ESTIMATES TO THE CORRECT TARGET POPULATION

In Section 2.3.5 it was pointed out that MAIC and STC as presently used, although based on the idea that the size of a relative treatment effect depends on the population, do not in general succeed in generating comparisons calibrated to the target population for the decision (unless the target population matches the AC trial population, which is unlikely). We propose that an additional assumption is made, which we call the *shared effect modifier assumption*, which will allow relative treatment effects to be projected into any population. One of the results of this assumption is that active-active treatment comparisons (e.g. B vs. C) may be transported into any target population, as any effect modifiers cancel out; indeed, the shared effect modifier assumption is required in order for this to be possible.

The shared effect modifier assumption applies to a set of active treatments $\mathcal{T}$, and states that (i) the effect modifiers of all treatments in $\mathcal{T}$ are the same, and (ii) the change in treatment effect caused by each effect modifier is the same for all treatments in $\mathcal{T}$.

This assumption is not required for MAIC or STC as currently used. However, if this assumption is deemed reasonable, then it may be leveraged to produce indirect comparisons in any given target population; we provide mathematical proof and examples in Appendix B. The shared effect modifier assumption is evaluated on a clinical and biological basis; treatments in the same class (i.e. sharing biological properties or mode of action) are likely to satisfy the shared effect modifier assumption, and those from different classes are not. In some circumstances, where effect modification is an artefact of the scale of measurement (possibly indicating a poor choice of scale), it will be valid for all active treatments. This assumption is, in fact, commonly made when meta-regression is used.[57] One of the reasons for assuming that treatments in the same class have the same effect modifiers, in the absence of overwhelming evidence to the contrary, is that relaxing this assumption could lead to seemingly perverse decisions. For example, it is not uncommon to switch from recommending no treatment to recommending a given treatment past a certain age, but it would be most unusual to switch among several treatments within the same class at various ages (say treatment B is most effective at age 50, treatment C at age 60, and treatment D at age 70, and so on). In the present "anchored" scenario, it is common that A is placebo or a standard treatment, and we might make the shared effect modifier assumption for the set of treatments $\mathcal{T} = \{B, C\}$.

The shared effect modifier assumption allows us to transpose indirect comparisons from any population where a relative effect has been observed, such as an AC trial, to any other population of interest P,

and recreate a full set of relative or absolute effects given an observed relative or absolute effect in the P population. In general, we make use of the following two relations concerning the marginal relative effects for a set of treatments $\mathcal{T}$ for which the shared effect modifier assumption holds:

$$d_{At(P)} - d_{At(Q)} = c \quad \forall t \in \mathcal{T}, \tag{13}$$

$$d_{tu(P)} = d_{tu(Q)} \quad \forall t, u \in \mathcal{T}. \tag{14}$$

(We assume here that A is not in $\mathcal{T}$, otherwise the situation is trivial.) That is, for any two populations P and Q, the difference in the relative A vs. t effects on the transformed scale is constant for all t in $\mathcal{T}$ (equation (13)), and relative t vs. u effects are constant across populations for any two active treatments t, u in $\mathcal{T}$ (equation (14)).

Therefore, if all relative effects are known in one population (say, the AC population) and for another population (say P) we are given an estimate of any single relative effect $d_{At(P)}$, where t is in $\mathcal{T}$, then immediately we can calculate estimates of all other relative effects $d_{Au(P)}$, where u is in $\mathcal{T}$, in the new population via equation (13) and/or (14). If we are given an estimate of a single absolute effect $\mu_{t(P)}$, where t is in $\mathcal{T}$, in the P population, then we can calculate estimates of all $\mu_{u(P)}$ absolute effects for all u in $\mathcal{T}$ via equation (14). Proofs are given in Appendix B, along with a step-by-step illustration of the calculations.

Equation (14) is of particular importance: if the shared effect modifier assumption holds for treatments B and C, then the estimated d_{BC} marginal relative treatment effect (whether obtained using anchored or unanchored MAIC/STC) will be applicable to *any* population.

3. MAIC AND STC APPLICATIONS IN THE LITERATURE

In the short time since the first papers on MAIC[3] and STC[6] were published, the use of these methods has increased dramatically – in particular MAIC, which has at least 10 published peer reviewed applications to date, along with numerous applications reported in conference abstracts. In this section we review the published applications of MAIC and STC in the literature, to examine how these new methods are being used in practice, and how well the methodology and assumptions underlying them are understood. Applied papers were found using a simple search amongst titles, abstracts, and keywords for "matching-adjusted indirect comparison" and "simulated treatment comparison" in Scopus and PubMed on 07/07/2016, by checking citing articles of the methodological papers,[3, 4, 6] and examining papers identified in a published scoping review.[58]

3.1 APPLICATIONS OF MAIC IN THE LITERATURE

In Table 1 we list the ten published applications of MAIC that our search identified in the literature to date, along with particular features and properties of the analyses, which we now discuss.

3.1.1 ANCHORED AND UNANCHORED COMPARISONS

The majority (60%) of the analyses involved randomised controlled trials with a common comparator. Of these, four out of six performed anchored indirect comparisons. Three out of six analyses involved an unanchored indirect comparison (one performed both anchored and unanchored indirect comparisons on different outcomes). In two of these, the unanchored approach was due to the outcome of interest being overall survival (OS) in a trial subject to treatment switches, where the placebo arm is contaminated by individuals crossing-over to active treatment after disease progression. The problem is avoided if progression free survival (PFS) rather than OS is the primary outcome (one analysis by Signorovitch *et al.*[59] performed an anchored indirect comparison for PFS and an unanchored indirect comparison for OS). An analysis by Sikirica *et al.*[60] had common placebo arms between the two trials, yet made an unanchored indirect comparison. The authors' justification was that, in the matching procedure, weights were additionally constrained to exactly balance placebo outcomes across trials. This method has yet to be evaluated either formally or through simulation studies, and its properties and performance in comparison with anchored methods are uncertain; in particular it is unlikely that balancing placebo outcomes is equivalent to relying on randomisation to remove residual differences due to unobserved prognostic variables.

A sizable proportion (40%) of analyses applied MAIC to single-arm trials, or in situations with no common comparator. The only choice in such a scenario is to perform an unanchored indirect comparison. As in all cases where unanchored indirect comparisons are performed, a strong assumption is made that all prognostic variables and all effect modifiers are accounted for and correctly specified –

an assumption largely considered to be implausibly strong. The published applications of unanchored MAIC acknowledge the possibility of residual bias due to unobserved prognostic variables and effect modifiers; however, it is not made clear that the accuracy of the resulting estimates is entirely unknown, because there is no analysis of the potential magnitude of residual bias, and hence no idea of the degree of error in unanchored MAIC estimates. Moreover, the inclusion of single-arm studies in an analysis is subject to the additional assumptions and biases incurred by these study designs.[61]

3.1.2 Availability of multiple studies for a treatment comparison

In half of the published analyses, issues arose with multiple IPD or aggregate populations for the same treatments. In both cases where multiple populations with IPD were available, the populations were simply pooled and treated as one large population. There was seemingly no attempt to account for the clustering of individuals within the component trials, which has been seen to incur bias and reduce power in the closely related context of IPD meta-analysis.[62] A better option in this scenario, in the absence of MAIC methodology which accounts for clustering, is to perform identical MAICs based on each IPD population, and then pool the relative effect estimates (on the linear predictor scale) with standard meta-analysis methods.[14, 63]

Multiple aggregate populations were pooled in two out of three cases, and analysed in separate MAICs in one other. When aggregate populations are pooled, this should always be done with relative effects on the linear predictor scale to avoid complications such as conflicts of scale (see Section 2.3.3). There are two equivalent ways in which such an analysis may be done: (i) perform identical MAICs into each AC population, and then pool the relative estimates $\hat{d}_{BC(AC)}$; or (ii) pool the aggregate AC populations and the relative estimates $\hat{d}_{AC(AC)}$, and then perform a single MAIC into the pooled population. In either case, the pooling of relative effect estimates should take place on the linear predictor scale using standard methods,[14, 63] and the resulting target population will be the (appropriately weighted) combination of the aggregate populations – which may or may not match the true target population for the decision.

3.1.3 Larger treatment networks

Two papers presented analyses involving more than three treatments, one by Signorovitch *et al.*[64] with four treatments arranged in a square network (Figure 1L) – essentially giving two possible common comparators (placebo and another active treatment) between the treatments of interest B and C, and another by Kirson *et al.*[65] with four treatments in a star network (Figure 1R), in this case having two competitor treatments C and D to make indirect comparisons with B.

Signorovitch *et al.* had access to IPD on the AB and BD studies, with aggregate data on the AC and CD studies; therefore two possible MAIC analyses could be performed, one via treatment A, and another via treatment D. The two resulting indirect comparison estimates are valid for different target populations – one for AC and one for CD – which were then pooled. The target population of the MAIC in this case is therefore a weighted combination of the AC and CD populations, which is unlikely to match the true target population for the decision.

Kirson *et al.* faced a similar scenario, where there were two competitor treatments C and D with aggregate AC and AD trial data with which to form an indirect comparison. Again, two MAICs were performed, this time giving an estimate of $d_{BC(AC)}$ and of $d_{BD(AD)}$. These relative estimates are not comparable as they are both valid for different target populations (AC and AD respectively), unless the two target populations have balanced distributions of effect modifiers. There is no way with current MAIC methods to achieve a coherent comparison of all four treatments in this case.

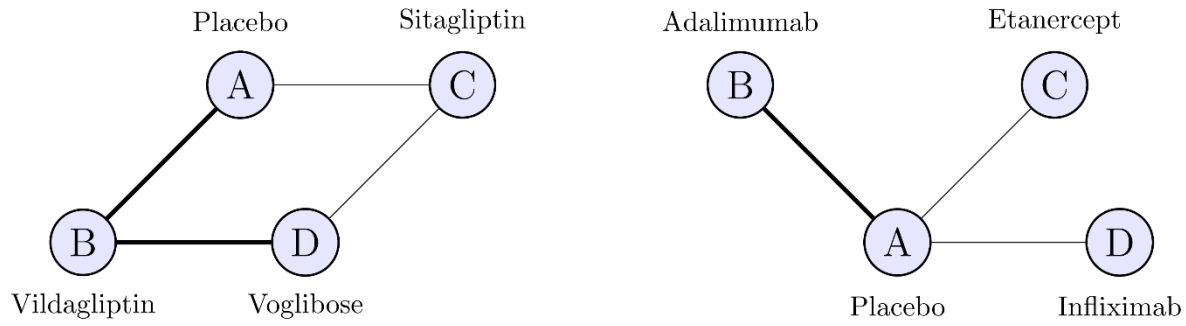

Figure 1: Network diagrams for analyses involving more than three treatments:
(L) Signorovitch et al.[64] *perform two MAICs via alternate common comparators; (R) Kirson* et al.[65] *perform two MAICs for two different competitor treatments. Thick edges indicate availability of IPD, thin edges indicate only aggregate data being available.*

3.1.4 Effective sample size and weight distributions

Only 40% of the published MAIC analyses made any mention of either effective sample size or the distribution of weights: three included an ESS, and one other included a summary of the distribution of weights. The reporting of at least one of these is fundamental to understanding and diagnosing poor overlap between the IPD and aggregate populations. When the ESS is markedly reduced, or equivalently the weights are highly variable, estimates become unstable and inferences depend heavily on just a small number of individuals. The three papers reporting ESS saw an 80% average reduction from the original sample size (range: 57–98%).

3.1.5 CHOICE OF MATCHING VARIABLES

The number of matching variables used in the published MAIC analyses varied between 2 and 17. Most analyses balanced the standard deviations of covariates as well as means or other summary statistics between the populations, but only one (Sikirica *et al.*[60]) included any interactions or higher order terms in the weighting model. The majority of published MAIC analyses therefore are subject to the additional assumptions set out in Section 2.3.4 due to the use of marginal covariate distributions instead of the joint distribution; in particular, an assumption must be made either regarding the balance of covariate correlations between populations, or regarding the lack of interaction terms in the implicit outcome model induced on the scale of the indirect comparison.

In no anchored analysis was there any attempt to justify the effect modifier status of the variables included in the weighting model, either with clinical expertise or with prior empirical evidence. The NICE Methods Guide[66] is explicit that effect modifier status should be justified prior to analysis. For unanchored comparisons, every prognostic variable as well as effect modifier should be included; only three analyses justified the included variables as being prognostic or effect modifying in any manner.

In general, published anchored MAIC analyses reported comparative estimates before and after the weighting adjustment, and noted any difference. However, the observation of a difference in relative effects after an analysis has been done should not be used to justify that an anchored MAIC should be preferred over a standard indirect comparison; such arguments amount to *post hoc* reasoning, whereas in the context of NICE technology appraisals all analyses should be clearly pre-specified.[66] No attempts were made prior to any analysis to assess the magnitude of impact of effect modifier imbalance on the indirect comparison (see Section 4.2.3).

In some cases where common placebo arms were present, placebo tests were performed as an attempt to justify the validity of the MAIC. However, such tests can only detect imbalance in observed or unobserved prognostic variables, and are completely unable to detect imbalances in observed or unobserved effect modifiers. It is arguable whether placebo tests in this context add any value at all: anchored indirect comparisons by design account for differences in prognostic variables between the two populations, so any imbalanced prognostic variables will not lead to bias in the indirect comparison but will cause a placebo test to "fail"; placebo tests should not be used to "justify" unanchored indirect comparisons due to their low power.

3.1.6 CHOICE OF SCALE

The choice of scale for an indirect comparison is important, as assumptions are implied on the indirect comparison scale regarding additivity of effects, definition of prognostic and effect modifying variables,

and distributional properties (see Section 2.3.3). Almost all published MAICs carried out the indirect comparison on the natural outcome scale. In many cases this led to indirect comparisons being made on scales not commonly used for meta-analyses, such as probability differences rather than log odds ratios. As in meta-analysis, the appropriate scale should be considered on a case-by-case basis, in light of the biological and clinical knowledge, with the default scale determined by existing literature.

Table 1: Applications of MAIC in the literature

Paper	Trials and treatments	AB sample size	AC sample size	Number of matching variables	Variables where evidence for effect modifier status is presented	Variables where evidence of imbalance is presented	Anchored or unanchored indirect comparison	Scale of outcome in indirect comparison	Usual scale of outcome
Signorovitch *et al.*[3]	*Company:* Adalimumab (B) vs. Placebo (A) Pooled two AB populations. *Competitor:* Etancercept (C) vs. Placebo (A)	Original: 1359 After excl. criteria: 1025 MAIC ESS: 591	330	10 (5 with SD)	2 – based on clinical reasoning.	4 (statistically significant)	Anchored	Response probability, percent change in PASI	logit (log OR), identity
Chang *et al.*[67]	*Company:* Bevacizumab + cisplatin (B) *Competitor:* Pemetrexed + cisplatin (C) Two single-arm trials.	Original: 2172 After excl. criteria: 72 MAIC ESS: 46	67	2 (0 with SD)	0 Variables described as "potentially prognostic".	2 (numerically different)	Unanchored	Median PFS	Identity
Signorovitch *et al.*[68]	*Company:* Nilotinib (B) vs. Imatinib (A) *Competitor:* Dasatinib (C) vs. Imatinib (A)	Original: A: 282 B: 283 After excl. criteria: A: 280 B: 273	A: 260 C: 259	10 (0 with SD)	0	3 (numerically different)	Unanchored	Proportion of MMR, PFS, and OS at 1 year	logit (log OR)

Paper	Trials and treatments	AB sample size	AC sample size	Number of matching variables	Variables where evidence for effect modifier status is presented	Variables where evidence of imbalance is presented	Anchored or unanchored indirect comparison	Scale of outcome in indirect comparison	Usual scale of outcome
Signorovitch *et al.*[64]	*Company:* Vildagliptin (B) vs. Placebo (A) Vildagliptin (B) vs. Voglibose (D) *Competitors:* Sitagliptin (C) vs. Placebo (A) Sitagliptin (C) vs. Voglibose (D) Two AC populations pooled for one analysis at one dose level.	Original: AB: 148 BD: 380 After excl. criteria: AB: 148 BD: 363	AC: 145 CD: 319	6 (5 with SD)	0 Noted large heterogeneity in previous meta-analyses.	3 (statistically significant)	Anchored Two MAICs performed with Placebo and Voglibose as common comparators, results then pooled.	Mean HbA1c	Identity
Kirson *et al.*[65]	*Company:* Adalimumab (B) vs. Placebo (A) *Competitors:* Etancercept (C) vs. Placebo (A) Infliximab (D) vs. Placebo (A)	Original: 313 After excl. criteria: 296 (for AC) 234 (for AD)	AC: 205 AD: 200	For AC: 12 (6 with SD) For AD: 17 (11 with SD)	0	2 for AC and 4 for AD (statistically significant)	Anchored	Response rates, percent change	logit (log OR), identity
Signorovitch *et al.*[59]	*Company:* Everolimus (B) vs. Placebo (A) *Competitor:* Sunitinib (C) vs. Placebo (A)	Original: 410 After excl. criteria: 394	171	9 (0 with SD)	0	3 (statistically significant)	Anchored for PFS Unanchored for OS	log hazard ratios	log hazard ratios

Paper	Trials and treatments	AB sample size	AC sample size	Number of matching variables	Variables where evidence for effect modifier status is presented	Variables where evidence of imbalance is presented	Anchored or unanchored indirect comparison	Scale of outcome in indirect comparison	Usual scale of outcome
Sikirica *et al.*[60]	*Company:* Guanfacine (B) vs. Placebo (A) Pooled two AB populations. *Competitor:* Atomoxetine (C) vs. Placebo (A)	Original: 631 After excl. criteria: A: 136 B: 82	A: 83 C: 84	4 (with SDs, pairwise interactions, quadratic and cubic terms)	0	1 (statistically significant)	Unanchored Weights are constrained such that placebo arms match exactly.	Mean ADHD scores	Identity
Sherman *et al.*[69]	*Company:* Everolimus (B) *Competitor:* Axitinib (C) No common comparator, other arms ignored.	Original: 277 After excl. criteria: 43	194	3	0 Variables found using latent class model as being influential on PFS.	3 (numerically different)	Unanchored	Median PFS	Identity
Van Sanden *et al.*[70]	*Company:* Simeprevir + peginterferon alfa 2a + ribavirin (B) *Competitor:* Peginterferon alfa 2a + ribavirin (C1-5) Single arms, multiple C populations.	Original: 107 After excl. criteria (MAIC ESS): For C1: 35 (29) For C2: 35 (15) For C3: 57 (14) For C4: 35 (26) For C5: 19 (17)	C1: 30 C2: 18 C3: 95 C4: 40 C5: 109	5-6	0 Consulted two experienced hepatologists for variables "relevant to treatment response".	Some numerical differences.	Unanchored	Proportion achieving sustained virologic response	logit (log OR)

Paper	Trials and treatments	AB sample size	AC sample size	Number of matching variables	Variables where evidence for effect modifier status is presented	Variables where evidence of imbalance is presented	Anchored or unanchored indirect comparison	Scale of outcome in indirect comparison	Usual scale of outcome
Swallow *et al.*[71]	*Company:* Daclatasvir + sofosbuvir (B) *Competitor:* Sofosbuvir + ribavirin (C) Pooled two C populations. All open label, single-arm.	Original: 153 After excl. criteria: 91	455	14 (3 with SD)	0	4 (statistically significant)	Unanchored	Proportion achieving sustained virologic response	logit (log OR)

3.2 APPLICATIONS OF STC IN THE LITERATURE

Our literature search returned only one published application of STC to date. Nixon *et al.*[72] present an analysis of oral therapies for the treatment of relapsing-remitting multiple sclerosis. A network diagram is shown in Figure 2. The AB population consisted of 1556 patients randomised to either fingolimod (B) or placebo (A) across two original trials with IPD. Unlike any MAIC analyses using pooled IPD, Nixon *et al.* correctly accounted for the clustering induced in the data by pooling across two study populations, by including a study-level baseline risk term in the outcome model (i.e. a separate intercept for each study). There were three trials with aggregate data: two comparing dimethyl fumarate (C) to placebo in a total of 2301 patients, and another comparing teriflunomide (D) to placebo in 1088 patients. Risk ratios and covariate distributions of the two AC trials were pooled simply using inverse variance weighting (essentially a fixed-effect meta-analysis of the two trials). Differences in covariate and outcome definitions between the AC and AD studies led Nixon *et al.* to produce two STC models, one using the AC definitions for prediction into the AC population, and the other using the AD definitions for prediction into the AD population.

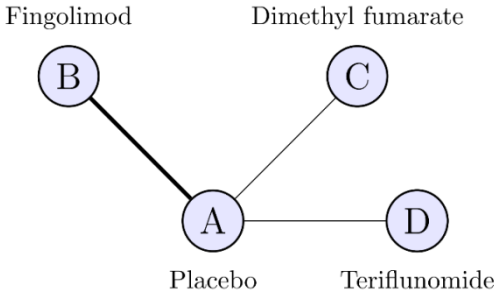

Figure 2: Network diagram for the STC analyses performed by Nixon *et al.*[72]
Nodes represent treatments, and edges between nodes represent studies comparing the corresponding treatments.

Of all published applications across MAIC and STC, Nixon *et al.*[72] are the only authors to attempt to justify effect modifier status of any variables; both expert clinical opinion and the results of previous subgroup analyses were used in evidence. There was no analysis of the imbalance in any covariates between the three populations beyond simple numerical differences, however the use of an AIC-based backwards selection algorithm to choose the final model suggests that the remaining covariates were significantly predictive of outcome. The outcome model itself was a linear probability model, using an identity link function to regress the probability of response against the covariates. As noted earlier this is an uncommon modelling choice, not least because such models can lead to predicted probabilities that lie outside the range 0 to 1. Similarly, this choice of model scale in this case also leads to problems with the anchored indirect comparison, which is constructed naturally on the (log) relative risk scale. It

therefore breaks the "anchoring" which is taken advantage of by the anchored indirect comparison. In the outcome regression, prognostic variables (and effect modifiers) are defined with respect to the linear probability scale, however the use of the log RR for the anchored indirect comparison means that prognostic variables *will not cancel*.

4. SUMMARY AND RECOMMENDATIONS

4.1 METHODOLOGICAL SUMMARY OF MAIC/STC IN RELATION TO EARLIER METHODS

4.1.1 OVERVIEW OF ASSUMPTIONS MADE BY DIFFERENT METHODS

MAIC and STC are both based upon methods of standardisation which date back several decades. In Section 2.1 we outlined the history and development of such methods, starting with model-based standardisation as an alternative to crude direct standardisation based on propensity score weighting or matching, and outcome regression models. The generalisation literature brought these methods into the context of generalising treatment effects to a target population, and described the assumptions necessary for such a process. The literature on treatment effect calibration utilised propensity score and outcome regression methods in the indirect comparison scenario with which we are concerned; however IPD were required in all studies. MAIC and STC extend these methods to deal with a lack of IPD, enabling the estimation of indirect comparisons with IPD on only one study (AB) and aggregate data on the other (AC). MAIC uses inverse propensity score weighting to form weighted mean estimators of the expected mean outcomes on treatments A and B in the AC population, where the propensity scores are found using a method of moments. STC estimates the mean outcomes by first fitting an outcome regression model to the IPD in the AB population, and then predicting outcomes for the AC population, if necessary by simulating individuals from the AC population.

As highlighted by the literature on generalisation (Section 2.1.4), identification of treatment effects relies on four core assumptions, regardless of the chosen methods with which the population adjustment or subsequent indirect comparison are to be made. These assumptions are summarised in Table 2. The first three of these assumptions are met by appropriately designed randomised studies. They are required by standard synthesis methods such as pair-wise meta-analysis and its extensions to indirect comparisons and network meta-analysis, and by MAIC/STC. In what follows we will assume that these three core assumptions have been met.

The fourth assumption is some form of constancy assumption, on an appropriate scale. The strength and scope of the constancy assumption varies depending on the method applied. A standard indirect comparison or network meta-analysis assumes constancy of relative effects on the linear predictor scale. Anchored forms of MAIC and STC rely on conditional constancy of relative effects, typically on the natural outcome scale. This means that the relative treatment effects are assumed constant between studies at any given level of the effect modifiers. No assumptions are needed regarding between-study

differences in the distribution of prognostic variables, because the first three assumptions guarantee balance within each study.

Unanchored MAIC and STC make the much stronger assumption of conditional constancy of absolute effects (called treatment-specific conditional constancy by Zhang *et al.*[37] in the calibration literature). This means that the absolute treatment effects are assumed constant at any given level of the effect modifiers and prognostic variables, and *all* effect modifiers and prognostic variables are required to be known. This is a far more demanding assumption, and it is widely accepted that it is very hard to meet. Unanchored comparisons based on disconnected networks and/or involving single-arm studies are therefore problematic.

4.1.2 THE IMPORTANCE OF SCALE AND ITS RELATION TO EFFECT MODIFICATION

Standard indirect comparison and network meta-analysis are carried out on a pre-specified scale, known as the linear predictor scale. This is typically the logit scale for proportions or the log scale for rate outcomes. When we refer to effect modifiers, we refer specifically to variables that modify treatment effects on the scale of the comparison (i.e. on the linear predictor scale for standard indirect comparison and network meta-analysis). MAIC and STC as currently practiced are typically carried out on the natural outcome scale, regardless of the conventional linear predictor scale, so that variables that are effect modifiers in standard indirect comparison might not be in MAIC/STC, and variables which are effect modifiers in MAIC/STC may not be effect modifiers in a standard indirect comparison analysis.

Although the identification of the "correct" scale for any specific outcome is debatable, there is a considerable literature (e.g. Deeks[73]) that shows that relative treatment effects for binary or rate outcomes are more stable across trials when they are expressed on logit or log scales, compared to absolute scales such as the risk difference, meaning there are fewer effect modifiers or that effect modification is weaker. Another concern of scale choice in the context of indirect comparisons is that different scales can lead to reverse conclusions, particularly for binary and rate outcomes when baseline event rates are diverse.[74] This reversal is due to the additivity assumption not being valid on all scales (indeed, it is impossible for additivity to hold on all scales).[75] The choice of an appropriate scale is therefore critical, and should be made using biological and clinical knowledge;[76] moreover, where a standard scale exists for a given outcome upon which additivity is commonly accepted, the use of an alternative scale is hard to justify.

In a decision making context, the possibility of effect modification has to be handled thoughtfully. The NICE Guide to the Methods of Technology Appraisal[66] is explicit that effect modifiers must be pre-specified and clinically plausible, and that supporting evidence must be provided from a thorough review of the subject area or from expert clinical opinion (see Section 5.2.7 of the NICE Methods

Guide). Moreover, although in the present context controlling for effect modifiers is undertaken to generate less biased population-average relative effects, the existence of an effect modifier can change the nature of the decision problem: for example if age is considered to be an effect modifier, it raises the possibility that a treatment that is effective at one age might not be effective at another.

For this reason, we make three related recommendations for how population-adjusted estimates should be obtained and presented. First, population adjustment, whether by propensity score weighting or by regression adjustment, should be performed with respect to the linear predictor scale usually employed in evidence synthesis for that outcome. Second, the propensity score weighting or regression adjustment should be applied to calibrate the relative-treatment effects and not to estimate individual absolute outcomes. Third, each variable used in population adjustment must be justified. This requires that (i) its status as an effect modifier needs to be supported by external quantitative evidence, expert opinion, or systematic review (as per the NICE Methods Guide[66]), and (ii) the degree of imbalance needs to be made explicit. These two factors should then be quantitatively combined to show the extent of bias reduction that is being achieved. This can be compared to the size of the unadjusted relative treatment effects obtained from a standard indirect comparison. Details of how this might be done are given below (Section 4.1.4).

One of the properties of the approach we are recommending is that, if there were no effect modifiers, no adjustment would occur even if we carried out propensity score weighting or regression adjustment: the estimates would be expected to be exactly those produced by standard indirect comparison and NMA methods. This is a desirable property for many reasons. First, anchored MAIC and STC methods as currently practiced represent a major departure from the models that are usually used. Second, as the methods stand, they open the prospect of different submissions adjusting for different variables, increasing the likelihood of inequitable and inconsistent decisions about different products for the same condition.

4.1.3 CALIBRATING POPULATION-ADJUSTED ESTIMATES TO THE CORRECT TARGET POPULATION

In Section 2.5 we proposed an additional assumption, called the *shared effect modifier assumption*, which is required to transport relative treatment effects into any population. The shared effect modifier assumption applies to a set of active treatments, and means that the effects of each treatment in this set are altered only by the same effect modifiers in the same way. In the present "anchored" AB and AC study scenario, it is common that A is placebo or a standard treatment, and we might make the shared effect modifier assumption for treatments B and C. This would then mean that there are no effect modifiers acting on the B vs. C comparison (since they all cancel out), and therefore the B vs. C estimate can be transported into any target population. The rationale for making the shared effect

modifier assumption is based on clinical and biological knowledge; the assumption will likely apply to treatments in the same class (i.e. sharing biological properties or mode of action).

4.1.4 UNANCHORED MAIC AND STC

Regulators are, increasingly, approving new products on the basis of single-arm studies, especially in oncology (50% of all FDA accelerated oncology approvals in 2015 were based on single-arm trials[77]), and reimbursement authorities are increasingly asked to assess treatments where only single-arm studies or disconnected networks are available. In this case unanchored MAIC or STC can be used to improve on "unadjusted" or naïve indirect comparisons by taking into account the different distributions of prognostic factors and effect modifiers in the two studies. (In the same way that MAIC and STC may improve upon standard "adjusted" indirect comparison by taking account of the distribution of effect modifiers.) However, it is essential that decision makers understand the different sources of error that attach to standard ("adjusted") indirect comparisons, naïve "unadjusted" indirect comparisons, and MAIC/STC in their anchored and unanchored forms. When non-randomised IPD are available on both studies, TSD 17 should be followed.[18]

In a standard adjusted indirect comparison, as long as there is no imbalance in effect modifiers, the only source of error in the relative effect estimates is the statistical sampling error, which depends on the sizes of the studies. If there is imbalance in effect modifiers this will cause an additional *systematic* error (bias). Population adjustment for those effect modifiers using propensity score weighting or regression adjustment will reduce this systematic error. Indeed, the systematic error will be eliminated if there are no further unobserved or uncontrolled effect modifiers. This is quite a strong assumption, but, given that standard indirect comparison assumes there are no effect modifiers in imbalance – whether observed or not – the assumption that there are no unobserved effect modifiers in imbalance represents a weaker assumption than standard indirect comparison, and seems a reasonable basis for a decision.

In the case of disconnected network or one-arm studies, the situation is quite different. A crude "unadjusted" indirect comparison will include sampling error plus systematic error due to the imbalance in both prognostic factors and effect modifiers. The size of this systematic error can certainly be reduced, and probably substantially, by appropriate use of MAIC or STC. Much of the literature on unanchored MAIC and STC acknowledges the possibility of residual bias due to unobserved prognostic variables and effect modifiers; however, it is not made clear that the accuracy of the resulting estimates is entirely unknown, because there is no analysis of the potential magnitude of residual bias, and hence no idea of the degree of error in the unanchored estimates. It is, of course, most unlikely that systematic error has been eliminated. Hoaglin,[78, 79] in a series of letters critiquing an unanchored comparison by Di Lorenzo *et al.*[80] based upon a matching approach similar to MAIC, remarked that, without providing

evidence that the adjustment compensates for the missing common comparator arms and the resulting systematic error, the ensuing results "are not worthy of consideration".

Therefore, if unanchored forms of population adjustment are to be presented, it is essential that submissions include information on the likely bias attached to the estimates, due to unobserved prognostic factors and effect modifiers distributed differently in the trials. The way in which residual systematic error is quantified is an area that requires further research. Some preliminary suggestions can be found in Appendix C.

4.1.5 Network meta-regression with limited IPD

In Section 2.2.3 we discussed a further class of methods based upon network meta-regression, in particular one derived from the hierarchical related regression introduced by Jackson *et al.*[47, 48] This approach differs conceptually from MAIC and STC, in that it models individual-level relationships and is able to provide internally consistent inferences at both the individual level and at an aggregate level like a standard indirect comparison. Methods such as MAIC and STC use IPD to predict average outcomes on study arms, and then produce an indirect comparison at the aggregate study level. We presented the general form of the model in equation (10). We regard this as a promising approach with some attractive properties. Most importantly: (i) it reduces to the gold-standard IPD network meta-regression if IPD are available for all trials, and (ii) it generalises naturally to connected networks of any size. This method requires much the same assumptions as our proposed forms of MAIC and STC; namely that all effect modifiers in imbalance are accounted for (conditional constancy of relative effects), and the shared effect modifier assumption. Although this method appears to represent a viable and attractive alternative to MAIC and STC, and their derivatives, we are not specifically recommending its use until the exact properties of this method, and its performance relative to methods such as MAIC and STC, has been investigated with thorough simulation studies.

4.1.6 Consistency across appraisals

The existing MAIC/STC literature has – quite appropriately – introduced methods for adjusting for differences in effect modifiers in anchored comparisons and both prognostic factors and effect modifiers in unanchored comparisons, using only limited IPD. There is a clear rationale for the use of such methods. However, our examination of the applied and methodological literature on MAIC/STC reveals that the ways in which these methods are being used represent new and unfamiliar models for relative treatment effect. Setting aside their failure to generate coherent population-adjusted estimates for the chosen target population, MAIC and STC also give very considerable leeway to investigators to choose anchored, or unanchored approaches, and to pick and choose variables to be adjusted for. Moreover, the existence of effect modifiers raises several issues which complicate the decision context, including

the possibility that different treatments might be optimal for different patients, and whether or not different treatments are affected by the same effect modifiers in the same way.

Following from this there is a high risk that the assumptions being made in one appraisal are fundamentally different from – even incompatible with – the assumptions being made a year later in another appraisal on the same condition. Therefore, in the interests of transparency and consistency, and to ensure equity for patients and a degree of certainty for those making submissions, it is essential to regularise how and under what circumstances these procedures should be used, and which additional analyses should be presented to support their use and assist interpretation.

We believe that the suggestions set out above in Sections 4.1.2-4 go a long way towards meeting these objectives. Some further proposals which have the same purpose are included in the recommendations below.

Table 2: Assumptions made by different methods for indirect comparisons.

	Method					
Assumptions made	**Standard indirect comparison, NMA**	**Network meta-regression***	**Unanchored MAIC**	**Anchored MAIC**	**Unanchored STC**	**Anchored STC**
Homogeneity of outcomes on each treatment	Y	Y	Y	Y	Y	Y
Stable unit treatment value	Y	Y	Y	Y	Y	Y
Within-study covariate balance (proper randomisation, ignorable treatment assignment)	Y	Y	N	Y	N	Y
Constancy						
Constancy of absolute effects	N	N	N	N	N	N
Conditional constancy of absolute effects	N	N	Y Typically on natural outcome scale.	N	Y Typically on natural outcome scale.	N
Constancy of relative effects	Y On linear predictor scale. For RE NMA relaxed to constancy in expectation.	N	N	N	N	N
Conditional constancy of relative effects	N	Y On linear predictor scale.	N	Y Typically on natural outcome scale.	N	Y Typically on natural outcome scale.
Shared effect modifiers	N/A	Y On linear predictor scale. Not required if IPD are available on both studies.	N†	N†	N†	N†

*The assumptions set out here are applicable to all forms of network meta-regression with varying combinations of IPD and aggregate data (both studies IPD, both studies aggregate data, one IPD and one aggregate), with the exception of the shared effect modifier assumption which is not required if IPD are available on both studies.

†The shared effect modifier assumption is not required, but may be additionally assumed in order to present estimates for another target population.

4.2 RECOMMENDATIONS FOR USE OF POPULATION-ADJUSTED INDIRECT COMPARISONS

The exact properties of population adjustment methodologies such as MAIC and STC, in anchored and unanchored forms, and their performance relative to standard indirect comparisons, can only be properly assessed by a comprehensive simulation exercise. For this reason we do not express any general preference for population reweighting or outcome regression. Similarly, we have not included the forms of network meta-regression that combine IPD and aggregate data[7, 47, 48] in our recommendations. These methods have attractive properties, but at this point there is no way telling how they would compare with MAIC or STC under failures in assumptions. Based on general principles and on the empirical findings presented in earlier sections, we can however draw some useful conclusions about the role of population-adjusted estimates of treatment effects, including the types proposed by MAIC and STC, in submissions to NICE.

These recommendations cover five areas:

1. The rationale for the use of population adjustment in submissions;
2. Justifying the use of population adjustment in both anchored and unanchored scenarios;
3. Variables for which population adjustment is required;
4. Generation of indirect comparisons for the appropriate target population;
5. Reporting guidelines for analyses involving population adjustment.

Appendix A provides flow charts summarising these recommendations, and describing the process of selecting a method for indirect comparison, undertaking the analysis, and presenting the results.

4.2.1 SCOPE OF POPULATION ADJUSTMENT METHODS

The rationale for employing population adjustment stems principally from two scenarios: (i) connected, comparative evidence is available, but standard synthesis methods are deemed inappropriate due to suspected effect modifiers in imbalance; (ii) no connected evidence is available, or comparisons are required involving single-arm studies. In either case, population-adjusted analyses must be fully justified following the criteria below. Population adjustment can only adjust for differences in observed covariate distributions between populations. Most notably, population adjustment cannot adjust for differences between trials relating to the treatments, such as treatment dosing formulation, treatment administration, co-treatments, treatment titration, or treatment switching.

4.2.2 ANCHORED VERSUS UNANCHORED FORMS OF POPULATION-ADJUSTED INDIRECT COMPARISON

The use of unanchored forms of population-adjusted indirect comparison requires that absolute outcomes can be reliably predicted into the aggregate AC trial. In practice, reliable prediction of this kind is very hard to obtain – it can only be achieved if the joint covariate set includes *every* prognostic variable and effect modifier acting in the AC trial. It is impossible to guarantee that all prognostic variables and effect modifiers are known or available, and therefore – by universal agreement: (i) randomized studies are required to infer the causal effects of treatment, rather than relying upon some form of covariate adjustment; and (ii) only *relative* treatment effects may be generalised from a trial, not the absolute outcomes.

For this reason we recommend that unanchored versions of population adjustment are avoided in situations where connected evidence is available (i.e. when a standard indirect comparison would be feasible). Only when anchored comparisons are not feasible, for example due to unconnected networks or comparisons involving single-arm trials, may unanchored comparisons be considered.

Recommendation 1: When connected evidence with a common comparator is available, a population-adjusted anchored indirect comparison may be considered. Unanchored indirect comparisons may only be considered in the absence of a connected network of randomised evidence, or where there are single-arm studies involved.

4.2.3 JUSTIFYING THE USE OF POPULATION-ADJUSTED ANCHORED INDIRECT COMPARISONS

Because the use of population adjustment itself makes a number of assumptions and complicates the process of treatment comparison in a connected network, evidence should be presented that population adjustment is likely to lead to superior estimates of treatment differences compared to standard methods.

Recommendation 2: Submissions using population-adjusted analyses in a connected network need to provide evidence that they are likely to produce less biased estimates of treatment differences than could be achieved through standard methods.

The argument for the use of population adjustment in a connected network is that (i) there are effect modifiers among the covariates on which data are available, that (ii) these effect modifiers are distributed differently in the AB (company) and AC (competitor) trials, and therefore (iii) that treatment effects estimated in the company's AB trial do not represent what would be expected in the (aggregate data) AC trial. To support the use of these methods in specific cases, Recommendation 2 therefore implies *two* forms of preliminary analysis. Specifically, is necessary to *both* establish that one

or more of the covariates is a known effect modifier, or can be plausibly considered as a potential effect modifier, *and* that these variables are in imbalance between the trials being considered.

4.2.3.1 Effect modifier status

Evidence that a variable is, or could be, an effect modifier for the outcome in question should therefore be presented. Such evidence could be based on external quantitative evidence, expert opinion, or systematic review (as per the NICE Methods Guide[66]). The concept of effect modification is scale dependent, and the relevant scale is the standard transformed scale used for the indirect comparison.

Recommendation 2a: Evidence must be presented that there are grounds for considering one or more variables as effect modifiers on the appropriate transformed scale. This can be empirical evidence, or an argument based on biological plausibility.

4.2.3.2 Evidence of substantial imbalance

Evidence should be brought forward that these specific effect modifiers are distributed differently in the AB and AC trials. A population-adjusted analysis should only be submitted if, putting together the magnitude of the supposed interaction with the extent of the imbalance, a material difference in the estimated treatment comparisons would be obtained. The case for controlling for a covariate needs to be presented in a quantitative way. For example, if the suspected effect modifier is represented as an interaction term of size γ, and the degree of imbalance between the AB and AC trials is $u = \bar{x}_{(AC)}^{EM} - \bar{x}_{(AB)}^{EM}$, the potential bias reduction compared to a standard indirect comparison will be γu. It needs to be shown that γu would represent a materially significant bias in relation to the observed treatment effects; the qualification of "substantial" bias should be considered in both a clinical (e.g. minimal clinically important difference) and statistical context. (If multiple effect modifiers are to be adjusted for, and if their joint distribution is available, then interaction terms may be taken into account to give a more accurate estimate of the potential overall bias reduction.)

Recommendation 2b: Quantitative evidence must be presented that population adjustment would have a material impact on relative effect estimates due to the removal of substantial bias.

4.2.4 JUSTIFYING THE USE OF POPULATION-ADJUSTED UNANCHORED INDIRECT COMPARISONS

In the scenario where a comparison is to be made using disconnected evidence or single-arm trials, an unanchored indirect comparison may be considered. The use of population adjustment in an unanchored indirect comparison requires that absolute outcomes can be reliably predicted. Those presenting such estimates should give evidence that the degree of bias due to imbalance in unaccounted for covariates

is acceptable, bearing in mind the size of the observed treatment effect. If this evidence cannot be provided or is limited, then any estimates or conclusions from the unanchored comparisons should be heavily caveated by noting: the amount of bias (systematic error) in these estimates is unknown, is likely to be substantial, and could even exceed the magnitude of treatment effects which are being estimated.

Recommendation 3: Submissions using population-adjusted analyses in an unconnected network need to provide evidence that absolute outcomes can be predicted with sufficient accuracy in relation to the relative treatment effects, and present an estimate of the likely range of residual systematic error in the "adjusted" unanchored comparison.

The manner in which this evidence is provided is likely to vary with the specific situation at hand, especially due to a likely lack of study evidence in the cases where population-adjusted unanchored indirect comparisons are suggested. We propose several potential avenues for quantifying the likely range of residual systematic error in Appendix C. Sensitivity analyses are advisable to assess how decisions are affected by a range of plausible biases in the effect estimates.

4.2.5 VARIABLES TO BE ADJUSTED FOR

The variables to be adjusted for in a population-adjusted analysis depend on whether an anchored or unanchored indirect comparison is to be formed.

For anchored indirect comparisons performed via population reweighting methods (e.g. MAIC), all effect modifiers, whether in imbalance or not, should be adjusted for to ensure balance and reduce bias. To avoid loss of precision due to over-matching, no prognostic variables which are not also effect modifiers should be adjusted for, as variables which are purely prognostic do not affect the estimated relative treatment effect.

For anchored indirect comparisons performed via outcome regression methods (e.g. STC), all effect modifiers in imbalance should be adjusted for, to reduce bias; further effect modifiers and prognostic variables may be adjusted for if this improves model fit (e.g. as measured by AIC or DIC). The inclusion of additional prognostic variables and effect modifiers can result in a gain in precision of the estimated treatment effect if the variable accounts for a substantial degree of variation in the outcome, but will not reduce bias any further.

For an unanchored indirect comparison, reliable predictions of absolute outcomes are required. Therefore, population adjustment methods should adjust for all effect modifiers and prognostic variables.

Recommendation 4: The following variables should be adjusted for in a population-adjusted analysis:

(a) For an anchored indirect comparison, propensity score weighting methods should adjust for all effect modifiers (in imbalance or not), but no prognostic variables. Outcome regression methods should adjust for all effect modifiers in imbalance, and any other prognostic variables and effect modifiers that improve model fit.

(b) For an unanchored indirect comparison, both propensity score weighting and outcome regression methods should adjust for all effect modifiers and prognostic variables, in order to reliably predict absolute outcomes.

4.2.6 SCALE OF INDIRECT COMPARISONS

In the absence of comprehensive simulation studies that might reveal the advantages and disadvantages of different methods in the circumstances of submissions to NICE, population-adjusted estimates should be generated in a way that is closely in line with general modelling practice, as expressed in the NICE Guide to the Methods of Technology Appraisal[66] and in ISPOR guidance.[81] To this end we recommend that methods are used which would yield the same results as standard methods in the case where there is no imbalance in effect modifiers.

Recommendation 5: Indirect comparisons should be carried out on the linear predictor scale, with the same link functions that are usually employed for those outcomes.

Accordingly, an anchored population adjustment of the AB treatment effect to estimate the relative BC effect in the AC population is formed as

$$\hat{\Delta}_{BC(AC)} = g\left(\bar{Y}_{C(AC)}\right) - g\left(\bar{Y}_{A(AC)}\right) - \left(g\left(\hat{Y}_{B(AC)}\right) - g\left(\hat{Y}_{A(AC)}\right)\right), \quad (15)$$

where $\bar{Y}_{t(AC)}$ is the observed summary outcome under treatment t in the AC trial, $\hat{Y}_{t(AC)}$ is the estimated summary outcome under treatment t in the AC trial, and $g(\cdot)$ is a suitable link function. Whichever population adjustment method the indirect comparison in (15) is reached by, an assumption must be made that all effect modifiers in imbalance are available and properly included in the analysis. If a link function is chosen that differs from the default as determined by existing literature for that outcome and condition, thorough justification must be given.

Similarly, an unanchored estimator is

$$\hat{\Delta}_{BC(AC)} = g\left(\overline{Y}_{C(AC)}\right) - g\left(\hat{Y}_{B(AC)}\right). \quad (16)$$

Whichever population adjustment method the indirect comparison in (16) is reached by, the assumption must be made that all effect modifiers and prognostic variables are available and properly accounted for.

4.2.7 Application of population adjustment to the appropriate target population

Population adjustment methods such as MAIC and STC as currently proposed are unable to achieve estimates in a target population other than that of the AC study. However, with the aid of an additional assumption (the shared effect modifier assumption) and by considering relative effects, we can take advantage of the mathematical relationships inherent in conditional consistency to derive estimates for the relevant target population (see Section 4.1.3).

Recommendation 6: The target population for any treatment comparison must be explicitly stated, and population-adjusted estimates of the relative treatment effects must be generated for this target population.

4.2.8 Reporting of population-adjusted analyses

When reporting population-adjusted analyses, the following themes should be considered and addressed explicitly:

1. The variables available in each study should be listed, along with their distributions (e.g. through box plots or histograms). Sufficient covariate overlap between the populations should be assessed: for population reweighting methods (such as MAIC), the number of individuals assigned zero weight should be reported; for outcome regression methods (such as STC), the amount of extrapolation required should be considered. For anchored comparisons this applies only to effect modifiers (see point 2); for unanchored comparisons all variables relevant to outcome should be presented.
2. Evidence for effect modifier status should be given (Section 4.2.3.1), along with the proposed size of the interaction effect and the imbalance between the study populations. The resulting potential bias reduction compared with a standard indirect comparison may be calculated by multiplying the interaction coefficient by the difference in means (see Section 4.2.3.2).
3. The distribution of weights should be presented for population weighting analyses, and used to highlight any issues with extreme or highly variable weights. Presentation of the effective sample

size may also be useful. ESS may be approximated using equation (7) – which is likely to be an underestimate – but provides clear warning where inferences are being made based on just a small number of individuals.

4. Measures of uncertainty, such as confidence intervals, should always be presented alongside any estimates. Care should be taken that uncertainty is appropriately propagated through to the final estimates (Section 2.4). For outcome regression methods, uncertainty is fully propagated for predictions into the aggregate population by the outcome regression model. For population reweighting methods, a robust sandwich estimator (as typical for MAIC) provides estimates of standard error which account for all sources of uncertainty. Other techniques include bootstrapping and Bayesian methods.
5. For an unanchored comparison, estimates of systematic error before and after population adjustment should be presented (Sections 4.1.4 and 4.2.4).
6. Present estimates for the appropriate target population using the shared effect modifier assumption if appropriate (Section 4.2.7), or comment on the representativeness of the aggregate population to the true target population.
7. In order to convey some clarity about the impact of any population adjustment, the standard indirect comparison estimate should be presented alongside the population-adjusted indirect comparison if an anchored comparison is formed; for an unanchored comparison, a crude unadjusted difference should be presented alongside the MAIC/STC estimate.

4.3 RESEARCH RECOMMENDATIONS

1. Development of new methods for population-adjusted treatment effects

MAIC and STC are methods for deriving population-adjusted average treatment effects that use IPD on one or more trials to estimate population average outcomes of treatments in other populations. The indirect comparison step is undertaken at the marginal (population average) level, as in a standard indirect comparison. The modified forms of MAIC and STC that we have recommended share this characteristic. Doubly robust methods, combining both propensity scores and outcome regression, are already established in the related literature, including that on calibration (Section 2.1.3). The advantage of doubly robust methods is that only one of the constituent models needs to be correct in order to provide valid estimates. However there has, to our knowledge, been no publication of doubly robust methods in the limited IPD scenario with which we are concerned (e.g. combining MAIC and STC methods into one doubly robust estimator). Another approach which has been seen to perform at least as well as traditional doubly robust estimators in other contexts is known as regression-adjusted matching, where regression adjustment is applied to propensity score matched data;[82] it is claimed that this approach reduces sensitivity to model specification. A similar approach could be used to combine MAIC and STC.

In Section 2.2.3 we drew attention to another class of methods based on network meta-regression with mixed IPD and aggregate data[7, 47, 48] which, in effect, combine the two levels of data by modelling the aggregate data as an integration over the IPD level data. This can be seen as a different class of model because the indirect comparison is also possible at the level of the conditional effects (at the individual level), as well as the marginal effects (at the aggregate level); we have referred to this different class of models as methods for *population-adjusted individual-level indirect comparisons*, as opposed to *population-adjusted study-level indirect comparisons*. Like the study-level methods, we suspect that these individual-level methods can be realised in several variants, and it would be of interest to explore these more fully.

2. Simulation studies

We have described the assumptions that must be made by population adjustment methods in order to achieve valid inference (Section 2.3). At present, it is entirely unknown how such methods might perform under varying degrees of failure in these assumptions. The priority must be that the properties of population adjustment methods in practical scenarios are probed through rigorous simulation studies, and the recommendations in this report reviewed and extended in the light of subsequent results.

3. Extent of error due to unaccounted for covariates

A robust and pragmatic approach is needed to quantify the possible extent of systematic error in unanchored indirect comparisons, the amount by which this systematic error is reduced by population adjustment, and therefore the residual systematic error inherent to the population-adjusted indirect comparison estimates. We provide initial suggestions of how this might be achieved in Appendix C, although further research is necessary to refine and validate these methods. This issue is of particular importance if comparative evidence based on single-arm studies or disconnected networks is to be seriously considered for the purposes of technology appraisal or guideline development. Methods are also needed for the assessment of error in anchored comparisons due to unaccounted for effect modifiers

4. Impact of availability of joint covariate distributions

At present, it is uncommon for published trials to report the full joint distribution of covariates. As such, population adjustment methods rely on additional assumptions in order to work with the reported marginal covariate distributions. The extent of error following the failure of these assumptions when working with marginal covariate distributions should be investigated through simulation studies; it is also likely that different population adjustment methods will perform differently in these scenarios. It would also be useful to obtain empirical data on the between-trial variation in the joint covariate distributions, to better inform population-adjusted analyses when only marginal covariate information is available, and to help understand the likely error in such scenarios.

5. Extension to larger networks

The scenario in which population adjustment methods such as MAIC and STC have been proposed is a "small network" scenario, where as few as two studies are available to inform an indirect comparison between two treatments. However, the motivation for and methodology underlying population adjustment methods is applicable to larger evidence bases, involving multiple treatments and several studies, which might typically be analysed using network meta-analysis. The extension of population adjustment into a larger network scenario with mixtures of IPD and aggregate data is therefore an area of interest, and should be compared with existing methods for network meta-regression.[7-10]

6. Uncertainty propagation

Full propagation of uncertainty through to the final estimates is important for informed decision-making. Formulating population adjustment in a Bayesian framework could be a convenient approach to fully accounting for all sources of variation, as well as enabling the inclusion of prior evidence into the models, and being readily integrated into a formal decision framework such as cost-effectiveness analysis. The properties of a Bayesian approach should be compared to current methods in simulation studies. A semi-Bayesian formulation of unanchored MAIC was previously proposed in a PhD thesis,[83] though we are yet to see any published applications of such an approach.

7. Software tools

Standardised computational tools for carrying out population adjustment, perhaps in the form of R packages (akin to GeMTC[84] for network meta-analysis) or code for Bayesian computation (e.g. for WinBUGS, JAGS, STAN), would help regularise the contents of submissions using these methods.

REFERENCES

1. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of Clinical Epidemiology* 1997;50:683-91.

2. Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence Synthesis for Decision Making 2: A Generalized Linear Modeling Framework for Pairwise and Network Meta-analysis of Randomized Controlled Trials. *Medical Decision Making* 2013;33:607-17.

3. Signorovitch JE, Wu EQ, Yu AP, Gerrits CM, Kantor E, Bao YJ, *et al.* Comparative Effectiveness Without Head-to-Head Trials A Method for Matching-Adjusted Indirect Comparisons Applied to Psoriasis Treatment with Adalimumab or Etanercept. *Pharmacoeconomics* 2010;28:935-45.

4. Ishak KJ, Proskorovsky I, Benedict A. Simulation and Matching-Based Approaches for Indirect Comparison of Treatments. *Pharmacoeconomics* 2015;33:537-49.

5. Signorovitch JE, Sikirica V, Erder MH, Xie JP, Lu M, Hodgkins PS, *et al.* Matching-Adjusted Indirect Comparisons: A New Tool for Timely Comparative Effectiveness Research. *Value Health* 2012;15:940-7.

6. Caro JJ, Ishak KJ. No Head-to-Head Trial? Simulate the Missing Arms. *Pharmacoeconomics* 2010;28:957-67.

7. Jansen JP. Network meta-analysis of individual and aggregate level data. *Research Synthesis Methods* 2012;3:177-90.

8. Saramago P, Sutton AJ, Cooper NJ, Manca A. Mixed treatment comparisons using aggregate and individual participant level data. *Statistics in Medicine* 2012;31:3516-36.

9. Donegan S, Williamson P, D'Alessandro U, Garner P, Smith CT. Combining individual patient data and aggregate data in mixed treatment comparison meta-analysis: Individual patient data may be beneficial if only for a subset of trials. *Statistics in Medicine* 2013;32:914-30.

10. Thom HH, Capkun G, Cerulli A, Nixon RM, Howard LS. Network meta-analysis combining individual patient and aggregate data from a mixture of study designs with an application to pulmonary arterial hypertension. *BMC Med Res Methodol* 2015;15:34.

11. Ades AE, Caldwell DM, Reken S, Welton NJ, Sutton AJ, Dias S. NICE DSU Technical Support Document 7: Evidence synthesis of treatment efficacy in decision making: a reviewer's checklist; 2012. http://www.nicedsu.org.uk

12. Dias S, Sutton AJ, Welton NJ, Ades AE. NICE DSU Technical Support Document 3: Heterogeneity: subgroups, meta-regression, bias and bias-adjustment; 2011. http://www.nicedsu.org.uk

13. Dias S, Sutton AJ, Welton NJ, Ades AE. NICE DSU Technical Support Document 6: Embedding evidence synthesis in probabilistic cost-effectiveness analysis: Software choices; 2011. http://www.nicedsu.org.uk

14. Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical Support Document 2: A generalised linear modelling framework for pair-wise and network meta-analysis of randomised controlled trials; 2011. http://www.nicedsu.org.uk

15. Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical Support Document 5: Evidence synthesis in the baseline natural history model; 2011. http://www.nicedsu.org.uk

16. Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical Support Document 1: Introduction to evidence synthesis for decision making; 2011. http://www.nicedsu.org.uk

17. Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. NICE DSU Technical Support Document 4: Inconsistency in networks of evidence based on randomised controlled trials; 2011. http://www.nicedsu.org.uk

18. Faria R, Hernandez Alava M, Manca A, Wailoo AJ. NICE DSU Technical Support Document 17: The use of observational data to inform estimates of treatment effectiveness for Technology Appraisal: Methods for comparative individual patient data.; 2015. http://www.nicedsu.org.uk

19. Glenny A, Altman D, Song F, Sakarovitch C, Deeks J. Indirect comparisons of competing interventions. *Health Technology Assessment* 2005;9:148.

20. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968;24:295-313.

21. Rosenbaum PR. Model-Based Direct Adjustment. *Journal of the American Statistical Association* 1987;82:387-94.

22. Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 1983;70:41-55.

23. Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society Series a-Statistics in Society* 2011;174:369-86.

24. Horvitz DG, Thompson DJ. A Generalization of Sampling without Replacement from a Finite Universe. *Journal of the American Statistical Association* 1952;47:663-85.

25. Kang JDY, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 2007;22:523-39.

26. Robins JM, Sued M, Lei-Gomez Q, Rotnitzky A. Comment: Performance of Double-Robust Estimators When "Inverse Probability" Weights Are Highly Variable. *Statistical Science* 2007; 10.1214/07-STS227D:544-59.

27. Funk MJ, Westreich D, Wiesen C, Sturmer T, Brookhart MA, Davidian M. Doubly Robust Estimation of Causal Effects. *American Journal of Epidemiology* 2011;173:761-7.

28. Bang H, Robins JM. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics* 2005;61:962-73.

29. Cole SR, Stuart EA. Generalizing Evidence From Randomized Clinical Trials to Target Populations. *American Journal of Epidemiology* 2010;172:107-15.

30. Stuart EA. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science* 2010;25:1-21.

31. Hartman E, Grieve R, Ramsahai R, Sekhon JS. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational

studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2015;178:757-78.

32. Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society Series a-Statistics in Society* 2008;171:481-502.

33. Hartman E, Hidalgo FD. What's the alternative?: An equivalence approach to balance and placebo tests. In: Department of Political Science UoCaB, Berkeley, ed.; 2011.

34. Zhang ZW. Covariate-Adjusted Putative Placebo Analysis in Active-Controlled Clinical Trials. *Statistics in Biopharmaceutical Research* 2009;1:279-90.

35. Nie L, Soon G. A covariate-adjustment regression model approach to noninferiority margin definition. *Stat Med* 2010;29:1107-13.

36. Nie L, Zhang ZW, Rubin D, Chu JX. Likelihood Reweighting Methods to Reduce Potential Bias in Noninferiority Trials Which Rely on Historical Data to Make Inference. *Annals of Applied Statistics* 2013;7:1796-813.

37. Zhang Z, Nie L, Soon G, Hu Z. New methods for treatment effect calibration, with applications to non-inferiority trials. *Biometrics* 2015; 10.1111/biom.12388.

38. Lu GB, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association* 2006;101:447-59.

39. White H. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* 1980;48:817-38.

40. Vartivarian S, Little RJ. Does weighting for nonresponse increase the variance of survey means? Paper presented at: JSM Proceedings, Survey Research Methods Section; Alexandria, VA. http://www.amstat.org/sections/srms/Proceedings/y2004/files/Jsm2004-000892.pdf

41. Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Statistics in Medicine* 2002;21:371-87.

42. Lambert PC, Sutton AJ, Abrams KR, Jones DR. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology* 2002;55:86-94.

43. Tudur Smith C, Williamson PR, Marson AG. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. *Statistics in Medicine* 2005;24:1307-19.

44. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *British Medical Journal* 2010;340.

45. Sutton AJ, Kendrick D, Coupland CAC. Meta-analysis of individual- and aggregate-level data. *Statistics in Medicine* 2008;27:651-69.

46. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd edn. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.

47. Jackson C, Best N, Richardson S. Improving ecological inference using individual-level data. *Statistics in Medicine* 2006;25:2136-59.

48. Jackson C, Best, Nicky, Richardson S. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2008;171:159-78.

49. Hainmueller J. Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis* 2012;20:25-46.

50. Belger M, Brnabic A, Kadziola Z, Petto H, Faries D. Inclusion of Multiple Studies in Matching Adjusted Indirect Comparisons (MAIC). Paper presented at: ISPOR 20th Annual Meeting; Philadelphia, PA, USA,

51. Belger M, Brnabic A, Kadziola Z, Petto H, Faries D. Alternative Weighting Approaches for Matching Adjusted Indirect Comparisons (MAIC). Paper presented at: ISPOR 20th Annual International Meeting; Philadelphia, PA, USA,

52. Brumback B, Berg A. On effect-measure modification: Relationships among changes in the relative risk, odds ratio, and risk difference. *Statistics in Medicine* 2008;27:3453-65.

53. Betts KA, Mittal M, Song J, Skup M, Joshi A. Relative efficacy of Adalimumab versus Secukinumab in active ankylosing spondylitis: a matching-adjusted indirect comparison. Paper presented at: European League Against Rheumatism; London, 8-11 Jun 2016

54. Maksymowych W, Strand V, Baeten D, Nash P, Thom H, Cure S, *et al.* Secukinumab for the treatment of ankylosing spondylitis: comparative effectiveness results versus Adalimumab using a matching-adjusted indirect comparison. Paper presented at: European League Against Rheumatism; London, 8-11 Jun 2016

55. Imbens GW, Rubin DB. Causal Inference for Statistics, Social, and Biomedical Sciences: Cambridge University Press; 2015.

56. Efron B. Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics* 1979;7:1-26.

57. Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu GB, Ades AE. Evidence Synthesis for Decision Making 4: Inconsistency in Networks of Evidence Based on Randomized Controlled Trials. *Medical Decision Making* 2013;33:641-56.

58. Veroniki AA, Straus SE, Soobiah C, Elliott MJ, Tricco AC. A scoping review of indirect comparison methods and applications using individual patient data. *BMC Med Res Methodol* 2016;16:1-14.

59. Signorovitch J, Swallow E, Kantor E, Wang X, Klimovsky J, Haas T, *et al.* Everolimus and sunitinib for advanced pancreatic neuroendocrine tumors: a matching-adjusted indirect comparison. *Experimental Hematology & Oncology* 2013;2:1-8.

60. Sikirica V, Findling RL, Signorovitch J, Erder MH, Dammerman R, Hodgkins P, *et al.* Comparative Efficacy of Guanfacine Extended Release Versus Atomoxetine for the Treatment of Attention-Deficit/Hyperactivity Disorder in Children and Adolescents: Applying Matching-Adjusted Indirect Comparison Methodology. *CNS Drugs* 2013;27:943-53.

61. Deeks J, Dinnes J, D'Amico R, Sowden A, Sakarovitch C. Evaluating non-randomised intervention studies. *Health Technology Assessment* 2003;7:186.

62. Abo-Zaid G, Guo B, Deeks JJ, Debray TPA, Steyerberg EW, Moons KGM, *et al.* Individual participant data meta-analyses should not ignore clustering. *Journal of Clinical Epidemiology* 2013;66:865-73.e4.

63. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0; 2011. http://handbook.cochrane.org

64. Signorovitch JE, Wu EQ, Swallow E, Kantor E, Fan L, Gruenberger J-B. Comparative Efficacy of Vildagliptin and Sitagliptin in Japanese Patients with Type 2 Diabetes Mellitus. *Clinical Drug Investigation* 2011;31:665-74.

65. Kirson NY, Rao S, Birnbaum HG, Kantor E, Wei RS, Cifaldi M. Matching-adjusted indirect comparison of adalimumab vs etanercept and infliximab for the treatment of psoriatic arthritis. *Journal of Medical Economics* 2013;16:479-89.

66. National Institute for Health and Care Excellence. Guide to the methods of technology appraisal 2013 no. PMG9. London; 2013. https://www.nice.org.uk/article/pmg9

67. Chang G-C, Ahn M-J, Wright E, Kim HT, Kim J-H, Kang JH, *et al.* Comparative effectiveness of bevacizumab plus cisplatin-based chemotherapy versus pemetrexed plus cisplatin treatment in East Asian non-squamous non-small cell lung cancer patients applying real-life outcomes. *Asia-Pacific Journal of Clinical Oncology* 2011;7:34-40.

68. Signorovitch JE, Wu EQ, Betts KA, Parikh K, Kantor E, Guo A, *et al.* Comparative efficacy of nilotinib and dasatinib in newly diagnosed chronic myeloid leukemia: a matching-adjusted indirect comparison of randomized trials. *Current Medical Research and Opinion* 2011;27:1263-71.

69. Sherman S, Amzal B, Calvo E, Wang X, Park J, Liu Z, *et al.* An Indirect Comparison of Everolimus Versus Axitinib in US Patients With Advanced Renal Cell Carcinoma in Whom Prior Sunitinib Therapy Failed. *Clinical Therapeutics* 2015;37:2552-9.

70. Van Sanden S, Pisini M, Duchesne I, Mehnert A, Belsey J. Indirect comparison of the antiviral efficacy of peginterferon alpha 2a plus ribavirin used with or without simeprevir in genotype 4 hepatitis C virus infection, where common comparator study arms are lacking: a special application of the matching adjusted indirect comparison methodology. *Current Medical Research and Opinion* 2016;32:147-54.

71. Swallow E, Song J, Yuan Y, Kalsekar A, Kelley C, Peeples M, *et al.* Daclatasvir and Sofosbuvir Versus Sofosbuvir and Ribavirin in Patients with Chronic Hepatitis C Coinfected with HIV: A Matching-adjusted Indirect Comparison. *Clinical Therapeutics* 2016;38:404-12.

72. Nixon R, Bergvall N, Tomic D, Sfikas N, Cutter G, Giovannoni G. No Evidence of Disease Activity: Indirect Comparisons of Oral Therapies for the Treatment of Relapsing–Remitting Multiple Sclerosis. *Advances in Therapy* 2014;31:1134-54.

73. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine* 2002;21:1575-600.

74. Norton EC, Miller MM, Wang JJ, Coyne K, Kleinman LC. Rank Reversal in Indirect Comparisons. *Value Health* 2012;15:1137-40.

75. van Valkenhoef G, Ades AE. Evidence Synthesis Assumes Additivity on the Scale of Measurement: Response to "Rank Reversal in Indirect Comparisons" by Norton et al. *Value Health* 2013;16:449-51.

76. Caldwell DM, Welton NJ, Dias S, Ades AE. Selecting the best scale for measuring treatment effect in a network meta-analysis: a case study in childhood nocturnal enuresis. *Research Synthesis Methods* 2012;3:126-41.

77. U.S. Food and Drug Administration. Hematology/Oncology (Cancer) Approvals & Safety Notifications. http://www.fda.gov/Drugs/InformationOnDrugs/ApprovedDrugs/ucm279174.htm (Accessed 10/08/2016).

78. Hoaglin DC, Cislo PR. An indirect comparison of everolimus versus sorafenib in metastatic renal cell carcinoma - a flawed analysis? *Expert Opinion on Pharmacotherapy* 2012;13:1077-8.

79. Hoaglin DC. An indirect comparison of everolimus versus sorafenib in metastatic renal cell carcinoma - a flawed analysis and a problematic response. *Expert Opinion on Pharmacotherapy* 2013;14:1705-6.

80. Di Lorenzo G, Casciano R, Malangone E, Buonerba C, Sherman S, Willet J, *et al.* An adjusted indirect comparison of everolimus and sorafenib therapy in sunitinib-refractory metastatic renal cell carcinoma patients using repeated matched samples. *Expert Opinion on Pharmacotherapy* 2011;12:1491-7.

81. Hoaglin DC, Hawkins N, Jansen JP, Scott DA, Itzler R, Cappelleri JC, *et al.* Conducting Indirect-Treatment-Comparison and Network-Meta-Analysis Studies: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: Part 2. *Value Health* 2011;14:429-37.

82. Kreif N, Grieve R, Radice R, Sekhon JS. Regression-adjusted matching and double-robust methods for estimating average treatment effects in health economic evaluation. *Health Services and Outcomes Research Methodology* 2013;13:174-202.

83. Han S. Statistical Methods for Aggregation of Indirect Information: Harvard University; 2014.

84. van Valkenhoef G, Lu G, de Brock B, Hillege H, Ades AE, Welton NJ. Automating network meta-analysis. *Research Synthesis Methods* 2012;3:285-99.

85. Hartung J, Knapp G. On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine* 2001;20:1771-82.

86. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine* 2001;20:3875-89.

87. Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. *Statistics in Medicine* 2002;21:3153-9.

88. Picard RR, Cook RD. Cross-Validation of Regression Models. *Journal of the American Statistical Association* 1984;79:575-83.

APPENDICES

Appendix A

A.1 PROCESS FOR POPULATION-ADJUSTED INDIRECT COMPARISONS

Figure 3: Flow chart for selecting methods for indirect comparisons

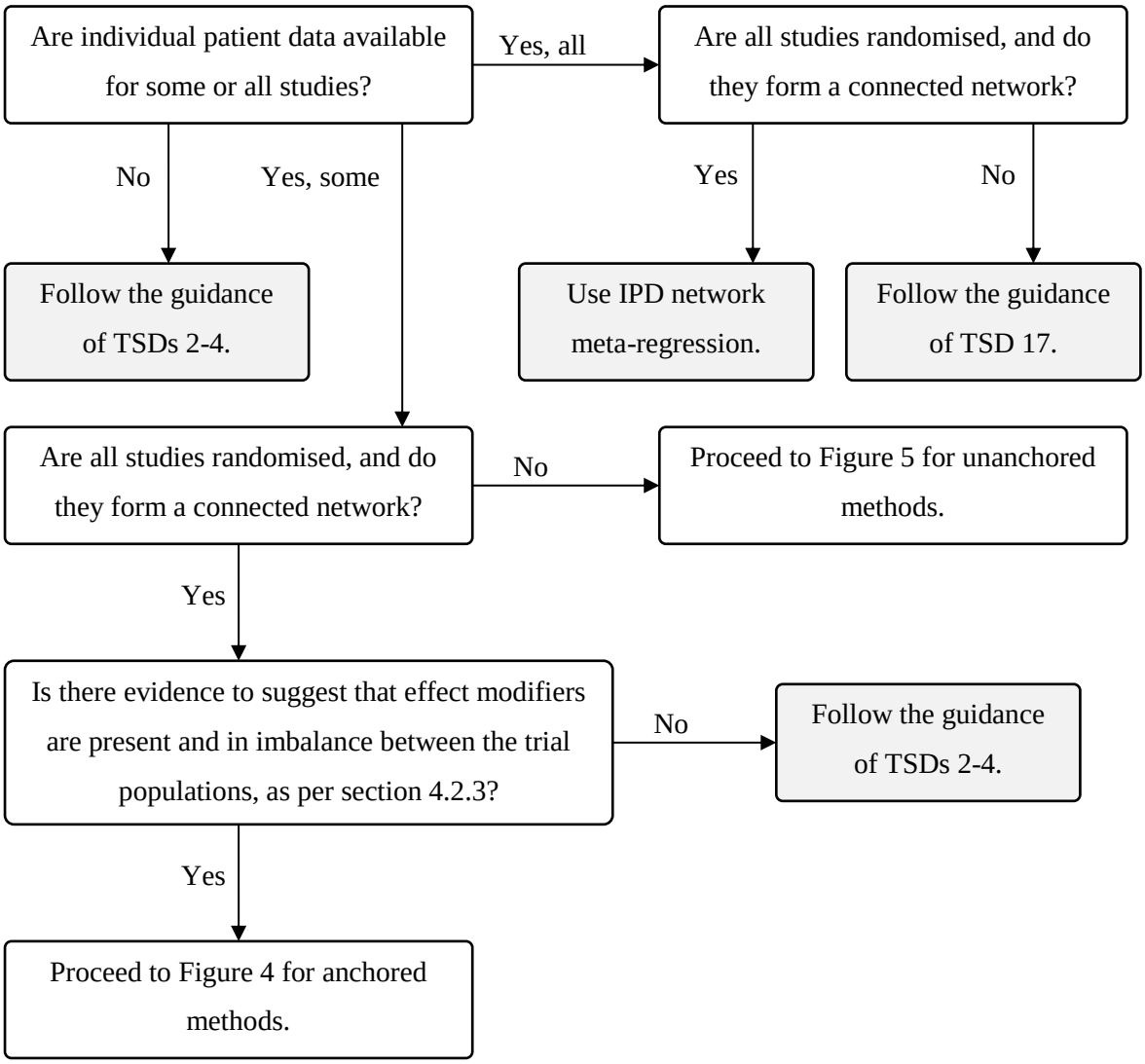

Figure 4: Anchored methods for population-adjusted indirect comparisons

Propensity score reweighting	**Outcome regression**
1. Provide evidence for effect modifier status on a suitable transformed scale (Section 4.2.3.1).	
2. Provide evidence that effect modifiers are in substantial imbalance between studies (Section 4.2.3.2).	
3a. Create a logistic propensity score model, which includes all effect modifiers but no prognostic variables. This is equivalent to a model on the log of the weights: $\log\left(w_{it}\right) = \alpha_0 + \boldsymbol{\alpha}_1^T \boldsymbol{X}_{it}^{EM}$	**3.** Fit an outcome model in the AB trial, which includes all effect modifiers in imbalance and any other prognostic variables or effect modifiers that improve model fit: $g\left(\mu_{t(AB)}\left(\boldsymbol{X}\right)\right) = \beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{X} + \left(\beta_B + \boldsymbol{\beta}_2^T \boldsymbol{X}^{EM}\right) I\left(t = B\right)$
3b. Estimate the weights using the method of moments to match effect modifier distributions between trials. This is equivalent to minimising $\sum_{t=A,B} \sum_{i=1}^{N_{t(AB)}} \exp\left(\boldsymbol{\alpha}_1^T \boldsymbol{X}_{it}^{EM}\right)$ when $\bar{\boldsymbol{X}}_{(AC)}^{EM} = \boldsymbol{0}$.	
4. Predict outcomes on treatments A and B in the AC trial by reweighting the outcomes of the AB individuals: $\hat{Y}_{t(AC)} = \frac{\sum_{i=1}^{N_{t(AB)}} Y_{it(AB)} \hat{w}_{it}}{\sum_{i=1}^{N_{t(AB)}} \hat{w}_{it}}$	**4.** Predict transformed outcomes on treatments A and B in the AC trial using the outcome model: $g\left(\hat{Y}_{t(AC)}\right) = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1^T \bar{\boldsymbol{X}}_{(AC)} + \left(\hat{\beta}_B + \hat{\boldsymbol{\beta}}_2^T \bar{\boldsymbol{X}}_{(AC)}^{EM}\right) I\left(t = B\right)$
5. Form the anchored indirect comparison in the AC population as: $\hat{\Delta}_{BC(AC)} = g\left(\bar{Y}_{C(AC)}\right) - g\left(\bar{Y}_{A(AC)}\right) - \left(g\left(\hat{Y}_{B(AC)}\right) - g\left(\hat{Y}_{A(AC)}\right)\right)$	
6. Calculate standard errors using a robust sandwich estimator, bootstrapping, or Bayesian techniques.	**6.** Calculate standard errors using the outcome model.
7. If justified, use the shared effect modifier assumption to transport the $\hat{\Delta}_{BC(AC)}$ estimate into the target population for the decision. Otherwise, comment on the representativeness of the AC population to the true target population.	
8. Present the distribution of estimated weights, and effective sample size.	**8.** Present standard model fit statistics.

Figure 5: Unanchored methods for population-adjusted indirect comparisons

Propensity score reweighting

1a. Create a logistic propensity score model, which includes all effect modifiers and prognostic variables. This is equivalent to a model on the log of the weights:

$$\log\left(w_i\right) = \alpha_0 + \boldsymbol{\alpha}_1^T \boldsymbol{X}_i$$

1b. Estimate the weights using the method of moments to match effect modifier distributions between trials. This is equivalent to minimising

$$\sum_{i=1}^{N_{B(B)}} \exp\left(\boldsymbol{\alpha}_1^T \boldsymbol{X}_i\right)$$

when $\bar{\boldsymbol{X}}_{(C)}^{EM} = \boldsymbol{0}$.

2. Predict outcomes on treatment B in the C trial by reweighting the outcomes of the B individuals:

$$\hat{Y}_{B(C)} = \frac{\sum_{i=1}^{N_{B(B)}} Y_{i(B)} \hat{w}_i}{\sum_{i=1}^{N_{B(B)}} \hat{w}_i}$$

Outcome regression

1. Fit an outcome model in the B trial, which includes all effect modifiers and prognostic variables:

$$g\left(\mu_{B(B)}\left(\boldsymbol{X}\right)\right) = \beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{X} + \left(\beta_B + \boldsymbol{\beta}_2^T \boldsymbol{X}^{EM}\right)$$

2. Predict transformed outcomes on treatments A and B in the AC trial using the outcome model:

$$g\left(\hat{Y}_{B(C)}\right) = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1^T \bar{\boldsymbol{X}}_{(C)} + \left(\hat{\beta}_B + \hat{\boldsymbol{\beta}}_2^T \bar{\boldsymbol{X}}_{(C)}^{EM}\right)$$

3. Form the unanchored indirect comparison in the C population as:

$$\hat{\Delta}_{BC(C)} = g\left(\bar{Y}_{C(C)}\right) - g\left(\hat{Y}_{B(C)}\right)$$

4. Calculate standard errors using a robust sandwich estimator, bootstrapping, or Bayesian techniques.

4. Calculate standard errors using the outcome model.

5. Provide evidence that absolute outcomes can be predicted with sufficient accuracy in relation to the relative treatment effects, and present an estimate of the likely range of residual systematic error (Section 4.2.4 and Appendix B). If this evidence cannot be provided or is limited, then state that the amount of bias in the indirect comparison is likely to be substantial, and could even exceed the magnitude of treatment effects which are being estimated.

6. If justified, use the shared effect modifier assumption to transport the $\hat{\Delta}_{BC(C)}$ estimate into the target population for the decision. Otherwise, comment on the representativeness of the C population to the true target population.

7. Present the distribution of estimated weights, and effective sample size.

7. Present standard model fit statistics.

Appendix B

B.1 TRANSPOSING INDIRECT COMPARISONS TO OTHER TARGET POPULATIONS

Under the assumption of shared effect modifiers for a set of treatments $\mathcal{T}$, we have the relations on the marginal relative treatment effects from equations (13) and (14):

Proposition 1: $d_{At(P)} - d_{At(Q)} = c \quad \forall t \in \mathcal{T}$

Proposition 2: $d_{tu(P)} = d_{tu(Q)} \quad \forall t,u \in \mathcal{T}$

which hold for any two populations P and Q.

Proof

Using additivity on an appropriate linear predictor scale, we write the transformed conditional absolute treatment effects $\eta_t(\boldsymbol{X},\boldsymbol{U})$ as

$$\eta_t(\boldsymbol{X},\boldsymbol{U}) = \beta_0 + \boldsymbol{\beta}_1^T\boldsymbol{X} + \boldsymbol{\varphi}_1^T\boldsymbol{U} + \left(\beta_t + \boldsymbol{\beta}_{2,t}^T\boldsymbol{X}^{EM} + \boldsymbol{\varphi}_{2,t}^T\boldsymbol{U}^{EM}\right) I(t \neq A), \tag{17}$$

where $\boldsymbol{X}$ and $\boldsymbol{U}$ are vectors of observed and unobserved covariates respectively (possibly with interactions or higher order terms), with corresponding subvectors of effect modifiers $\boldsymbol{X}^{EM}$ and $\boldsymbol{U}^{EM}$. Equation (17) represents the underlying (transformed) outcome model, which cannot be estimated directly as $\boldsymbol{U}$ are unobserved.

Using the shared effect modifier assumption on the set of treatments $\mathcal{T}$, which means that $\boldsymbol{\beta}_{2,t} = \boldsymbol{\beta}_2$ and $\boldsymbol{\varphi}_{2,t} = \boldsymbol{\varphi}_2 \ \forall t \in \mathcal{T}$ we rewrite the outcome model (17) for $t \in \mathcal{T}$ as

$$\eta_t(\boldsymbol{X},\boldsymbol{U}) = \beta_0 + \boldsymbol{\beta}_1^T\boldsymbol{X} + \boldsymbol{\varphi}_1^T\boldsymbol{U} + \left(\beta_t + \boldsymbol{\beta}_2^T\boldsymbol{X}^{EM} + \boldsymbol{\varphi}_2^T\boldsymbol{U}^{EM}\right) I(t \neq A). \tag{18}$$

We are now ready to proceed in proving the two propositions.

Firstly to prove proposition 1 see that, for any treatment $t \in \mathcal{T}$ and any two populations P and Q, we can write the marginal relative effects in terms of the conditional absolute effects by taking expectation over the population P and using equation (18):

$$
\begin{aligned}
d_{At(P)} &= \mathbb{E}_{(P)}\left(\eta_t\left(\boldsymbol{X},\boldsymbol{U}\right)-\eta_A\left(\boldsymbol{X},\boldsymbol{U}\right)\right) \\
&= \beta_t + \mathbb{E}_{(P)}\left(\boldsymbol{\beta}_2^T \boldsymbol{X}^{EM} + \boldsymbol{\varphi}_2^T \boldsymbol{U}^{EM}\right)
\end{aligned}
$$

$$
\begin{aligned}
d_{At(Q)} &= \mathbb{E}_{(Q)}\left(\eta_t\left(\boldsymbol{X},\boldsymbol{U}\right)-\eta_A\left(\boldsymbol{X},\boldsymbol{U}\right)\right) \\
&= \beta_t + \mathbb{E}_{(Q)}\left(\boldsymbol{\beta}_2^T \boldsymbol{X}^{EM} + \boldsymbol{\varphi}_2^T \boldsymbol{U}^{EM}\right)
\end{aligned}
$$

The difference in the marginal relative effects is then

$$
d_{At(P)} - d_{At(Q)} = \mathbb{E}_{(P)}\left(\boldsymbol{\beta}_2^T \boldsymbol{X}^{EM} + \boldsymbol{\varphi}_2^T \boldsymbol{U}^{EM}\right) - \mathbb{E}_{(Q)}\left(\boldsymbol{\beta}_2^T \boldsymbol{X}^{EM} + \boldsymbol{\varphi}_2^T \boldsymbol{U}^{EM}\right). \tag{19}
$$

Note that the RHS of (19) does not depend on t. Therefore, for any two populations P and Q, $d_{At(P)} - d_{At(Q)} = c$ holds for all $t \in \mathcal{T}$.

To prove proposition 2 we proceed similarly, writing the marginal relative effects between any two treatments $t,u \in \mathcal{T}$ in any two populations P and Q as

$$
\begin{aligned}
d_{tu(P)} &= \mathbb{E}_{(P)}\left(\eta_u\left(\boldsymbol{X},\boldsymbol{U}\right)-\eta_t\left(\boldsymbol{X},\boldsymbol{U}\right)\right) \\
&= \beta_u - \beta_t + \mathbb{E}_{(P)}\left(\boldsymbol{\beta}_2^T \boldsymbol{X}^{EM} + \boldsymbol{\varphi}_2^T \boldsymbol{U}^{EM}\right) - \mathbb{E}_{(P)}\left(\boldsymbol{\beta}_2^T \boldsymbol{X}^{EM} + \boldsymbol{\varphi}_2^T \boldsymbol{U}^{EM}\right) \\
&= \beta_u - \beta_t
\end{aligned}
$$

$$
\begin{aligned}
d_{tu(Q)} &= \mathbb{E}_{(Q)}\left(\eta_u\left(\boldsymbol{X},\boldsymbol{U}\right)-\eta_t\left(\boldsymbol{X},\boldsymbol{U}\right)\right) \\
&= \beta_u - \beta_t + \mathbb{E}_{(Q)}\left(\boldsymbol{\beta}_2^T \boldsymbol{X}^{EM} + \boldsymbol{\varphi}_2^T \boldsymbol{U}^{EM}\right) - \mathbb{E}_{(Q)}\left(\boldsymbol{\beta}_2^T \boldsymbol{X}^{EM} + \boldsymbol{\varphi}_2^T \boldsymbol{U}^{EM}\right) \\
&= \beta_u - \beta_t
\end{aligned}
$$

Therefore $d_{tu(P)} = d_{tu(Q)}$ holds for all $t,u \in \mathcal{T}$.

B.2 EXAMPLE

To see the application of these propositions in practice, consider an example where the following log odds ratios in the AC population have been estimated to be:

$$
\hat{d}_{AB(AC)} = 1.3 \qquad \hat{d}_{AC(AC)} = 0.8.
$$

Furthermore, in a population P the log odds ratio for treatment B compared to A is estimated to be $\hat{d}_{AB(P)} = 0.7$. We make the shared effect modifier assumption for treatments $\{B,C\} = \mathcal{T}$.

Using proposition 1 we have that $c = \hat{d}_{AB(P)} - \hat{d}_{AB(AC)} = -0.6$, and the log odds ratio for treatment C compared to A can be inferred to be $\hat{d}_{AC(P)} = \hat{d}_{AC(AC)} + c = 0.2$.

Alternatively, from the AC trial we have that $\hat{d}_{BC(AC)} = \hat{d}_{AC(AC)} - \hat{d}_{AB(AC)} = -0.5$. Now using proposition 2, we have that $\hat{d}_{BC(AC)} = \hat{d}_{BC(P)}$, and the log odds ratio for treatment C compared to A is again inferred to be $\hat{d}_{AC(P)} = \hat{d}_{AB(P)} + \hat{d}_{BC(P)} = 0.2$.

Appendix C

C.1 QUANTIFYING SYSTEMATIC ERROR IN UNANCHORED INDIRECT COMPARISONS

Unanchored indirect comparisons are not protected from imbalances in prognostic variables, unlike anchored indirect comparisons, as they do not rely on within-study randomisation. Unanchored indirect comparisons are therefore susceptible to large amounts of systematic error unless all prognostic variables and effect modifiers are accounted for. It is therefore necessary to attempt to quantify the possible extent of any residual systematic error resulting from unobserved prognostic variables and effect modifiers. The simplest way to quantify residual systematic error is by comparing observed and predicted outcomes on the company's treatment B in a range of different studies in the target population, however this might not be a viable option if there are no such studies available. It should be noted however that unobserved covariates are only one source of heterogeneity between studies (and bias in an ensuing indirect comparison); for example, differences in study design and conduct will also introduce heterogeneity, but cannot be accounted for with methods such as MAIC or STC. The way in which residual systematic error is quantified is therefore an area that requires further research. We present some initial suggestions here.

C.1.1 OUT-OF-SAMPLE METHODS

Firstly, the possible extent of systematic bias present in a crude "unadjusted" indirect comparison can often be quantified as follows. First identify a set of external studies in the target population with aggregate data on the relevant outcome. Then carry out a random-effects pooling across absolute outcomes on study arms in the target population, controlling for treatment. Usually one finds that the between-studies standard deviation τ in absolute outcomes far exceeds the AB relative treatment effect, let alone the effect of an active treatment B against an active competitor C, whether in the same class or not. Predicted outcomes on treatment B in each of the study arms used in the pooling can be obtained using MAIC or STC, and a similar pooling performed. If all prognostic variables and effect modifiers are accounted for, then the between-studies variation of the predicted outcomes, say τ_*^2, will match that of the observed outcomes τ^2 (that is, residual variation will be minimised). Conversely, lower between-studies variation of predicted outcomes would be expected if some prognostic variables and/or effect modifiers remain unaccounted for. The ratio of the between-studies variance in predicted to observed outcomes, τ_*^2 / τ^2, could be interpreted as the proportion of systematic error "explained" by the included covariates. It is likely that, in practice, limited study data will be available, and therefore the estimation of between-studies variance may be difficult. Robust frequentist

methods such as the Hartung-Knapp-Sidik-Jonkman estimator,[85-87] or Bayesian methods using the same plausibly vague prior distributions for τ^2 and τ_*^2 may be appropriate here.

Methods based on between-studies variance should be underpinned by a protocol-driven systematic review to prevent selection of an overly homogeneous sample of studies for inclusion. Likewise, the company's own trials would be expected to be more homogeneous than a wider selection of trials in the target population.

C.1.2 IN-SAMPLE METHODS

Other approaches for quantifying the systematic error in unanchored comparisons are possible. For example, if STC is used, cross-validation methods (e.g. Picard and Cook[88]) enable the estimation of the prediction error in the outcome model, which is largely due to missing prognostic variables and/or effect modifiers. k -fold cross-validation is a frequently used method, in which the IPD are split into k equal-sized sets. Each of the k sets is omitted in turn at the model fitting stage, and used as a validation set to check the model predictions. Prediction error may then be averaged over the k sets. A value of $k=10$ is often used, although the choice of k should be based on the situation at hand, particularly with reference to the available sample size, as there is a bias-variance trade-off. If outcome regression is used, R^2 values may be used to assess the predictive performance of the model; R^2 may be interpreted as the proportion of variance explained by the model, similarly to the between-studies variance ratio described above. A general disadvantage in our context of cross-validation, R^2, and other "in-sample" methods for checking predictive accuracy (as opposed to the "out-of-sample" methods above), is that the individuals in the IPD trial are likely to be more homogeneous than those of the wider target population, thus leading to overconfidence in the abilities of STC to predict outcomes in the target population. In-sample methods in general will most likely underestimate the true amount of residual variation.

Methods for estimating the likely error in unanchored population-adjusted indirect comparisons are a key area for further research.