NICE DSU TECHNICAL SUPPORT DOCUMENT 17:

THE USE OF OBSERVATIONAL DATA TO INFORM ESTIMATES OF TREATMENT EFFECTIVENESS IN TECHNOLOGY APPRAISAL: METHODS FOR COMPARATIVE INDIVIDUAL PATIENT DATA

REPORT BY THE DECISION SUPPORT UNIT

May 2015

Rita Faria,[1*] Monica Hernandez Alava,[2*] Andrea Manca,[1] Allan J Wailoo[2]

[1] Centre for Health Economics, University of York

[2] School of Health and Related Research, University of Sheffield

* Co-first authors

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 manufacturers, 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 Sarah Garner, Richard Grieve, Martin Hoyle, Noemi Kreif, Dan Mullins and the NICE team led by Melinda Goodall, 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:

Faria, R., Hernandez Alava, M., Manca, A., Wailoo, A.J. 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.
Available from http://www.nicedsu.org.uk

EXECUTIVE SUMMARY

Non-randomised studies may be used either to complement the evidence base represented by randomised controlled trials (RCTs) or as the source of evidence for a specific effectiveness parameter if randomised data are not available. The current 2013 NICE Guide to the Methods of Technology Appraisal (Methods Guide) recognises the potential biases that may arise from the use of non-RCT data, namely from confounding, lack of blinding, incomplete follow-up and lack of a pre-specified end-point. It recommends that potential biases should be identified, and ideally quantified and adjusted for. However, it does not provide guidance on how to estimate treatment effect or the appropriate methods to deal with potential biases. Lack of clear guidance on the use of evidence from non-RCTs may lead to inappropriate and inconsistent use of methods which in turn lead to biased estimates which may have potential adverse consequences for decisions on the effectiveness and cost-effectiveness of health technologies.

The objectives of this Technical Support Document (TSD) are to (i) summarise commonly available methods to analyse comparative individual patient data (IPD) from non-RCTs to obtain estimates of treatment effect to inform NICE Technology Appraisals (TAs) and (ii) to propose a set of recommendations to improve the quality and transparency of future assessments. It includes a summary of:

- The most commonly used methods for non-randomised IPD (Section 2);
- TAs which used non-randomised data to inform estimates of treatment effect for the cost-effectiveness analysis (Section 3);

It also includes

- An algorithm to aid selection of the appropriate method(s) for the analysis (Section 4.1);
- A review of existing checklists for quality assessment of the analysis of non-randomised studies (Section 4.2.1);
- A novel checklist (the QuEENS checklist) to assess the quality of the analysis of non-randomised studies (Section 4.2.2); and
- A summary of findings and final recommendations (Sections 5.1-5.2).

This TSD provides practical guidance on the methods that are relatively straightforward to apply and are most commonly used in statistical and econometric analysis of non-randomised

data. Specifically, it focuses on approaches that can be applied using standard statistical software without additional bespoke programming and advanced econometric/statistical skills. It is therefore aimed at those typically engaging in the NICE TA process whether submitting or reviewing evidence. The reviews and tools presented in this TSD are intended to help improve the quality of the analysis, reporting, critical appraisal and interpretation of estimates of treatment effect from non-RCT studies.

CONTENTS

TABLES AND FIGURES

ABBREVIATIONS

ATE	Average Treatment Effect
ATT	Average Treatment effect on the Treated
DiD	Difference-in-Differences
ERG	Evidence Review Group
FAD	Final Appraisal Document
GRACE	Good ReseArch for Comparative Effectiveness
IPD	Individual Patient Data
IPW	Inverse Probability Weighting
ISPOR	International Society for Pharmacoeconomics and Outcomes Research
IV	Instrumental Variable
LATE	Local Average Treatment Effect
MTE	Marginal Treatment Effect
NICE	National Institute for Health and Care Excellence
OLS	Ordinary Linear Regression
QuEENS	Quality of Effectiveness Estimates from Non-randomised Studies
RA	Regression Adjustment
RCT	Randomised Controlled Trial
RD	Regression Discontinuity
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
TA	Technology Appraisal
TSD	Technical Support Document

GLOSSARY

Term	Definition
Average treatment effect (ATE)	A treatment (or policy) effect averaged across the population
Average treatment effect on the treated (ATT)	The average treatment effect for the subgroup of individuals who have received the treatment
Caliper	A distance, a radius
Conditional independence assumption	See Ignorability of treatment assumption
Consistent Estimator	An estimator which converges (in probability) to the true population value as the sample size increases
Efficient Estimator	The "best" possible estimator amongst those with the same properties. "best" can be defined in different ways, one commonly used definition is "smallest variance"
Endogenous explanatory variable	An explanatory variable that is correlated with the error term. The correlation can arise because of omitted variables, measurement error or simultaneity
Endogeneity	A term used to describe the presence of an endogenous explanatory variable
Estimator	A rule for combining data to derive the value for the parameter of interest in the population
Exclusion restriction	A restriction which assumes that a variable does not appear in a model or in one equation in a (multiple equation) model
Exogenous variable	Any variable uncorrelated with the error term in the model under consideration
Homogeneous treatment effect	All individuals under consideration have the same treatment effect
Identification	The population parameter of interest can be consistently estimated using the data available
Ignorability of treatment assumption	Conditional on a chosen set of observed covariates, the potential outcomes are independent of the treatment assignment.
Local average treatment effect (LATE)	An average treatment effect which does not apply to the whole population but only to a small subpopulation of individuals
Monotonic	Always non-decreasing or always non-increasing
Non-parametric methods	Methods which do not rely on the estimation of distributions through their associated parameters
Parametric methods	Methods which rely on distributional assumption
Selection bias	Bias that arises when comparing the effect of a treatment in groups that are systematically different on variables that have an independent effect on the outcome on interest
Treatment effect	The change in outcome attributable to the treatment explanatory variable
Unbiased estimator	An estimator whose mean (of its sampling distribution) equals the population value
Unconfoundeness assumption	See Ignorability of treatment assumption

1. INTRODUCTION

1.1. BACKGROUND AND MOTIVATION

Non-randomised studies may be used as a complement to randomised controlled trials (RCTs) or as the single source of evidence for a specific parameter if randomised data are not available. In RCTs, random allocation of the study units (e.g. patients, clusters) between intervention(s) and comparator(s) ensures that (observable and unobservable) factors that can influence the outcome(s) of interest are evenly balanced between treatment groups. In its simplest form, an RCT evaluates the effect of an intervention (the 'treatment') relative to the comparator, such as no treatment, a placebo or another intervention, by comparing the outcomes of the individuals randomised to the intervention (the 'treated group') with those randomised to the comparator (the 'control group'). Since randomisation is designed to ensure that the factors affecting outcomes are evenly balanced between treatment groups, the change in outcome attributable to the treatment ('the treatment effect') in the study population is the difference in outcomes between the treated and the control groups.

In non-randomised studies, treatment is assigned based on a mechanism other than random allocation. Deeks *et al.*[1] proposes a taxonomy of study designs but others are also available. Non-randomised studies can be differentiated by whether: (i) treatment allocation is under the control of the investigator or outside their control, (ii) there is a comparison (control) group and (iii) data collection was pre-planned (prospective) or ad-hoc (retrospective). Non-controlled studies lack a comparison group, which means that inferences on the treatment effect must rely on before-and-after comparisons of the outcome of interest. Quasi-randomised trials are studies in which individuals are allocated to intervention and control groups by the investigator, but the method of allocation falls short of genuine randomisation (e.g. allocation by month of birth or by hospital record number). In quasi-experimental trials, the investigator has control over the allocation of the individuals to the treatment groups but does not attempt randomisation (e.g. allocation by patient or physician preference). In before-and-after studies, the outcomes are compared before and after treatment receipt; however, before-and-after studies may also include a control group which can allow for a comparison of the mean change in outcome between the control and the treatment group. A concurrent cohort study follows individuals who received different interventions over time. Cohort studies can be prospective, if the data collection was pre-planned, or retrospective. In

historical cohort studies, the treated cohort is compared with an untreated cohort who did not receive the intervention in a previous period, i.e. the individuals are not studied concurrently. A case-control study compares the exposure to treatment and outcomes of individuals with and without a given outcome (cases and controls, respectively). Cross-sectional studies examine the relationship between the outcome and other variables of interest at one point in time. In a case series, the outcome of a number of individuals who received the intervention is described.

Treatment effect estimates derived from non-randomised studies are at greater risk of bias. Selection bias occurs when the individuals in the treatment and control groups are systematically different in prognostic factors. Selection bias is more likely to occur in non-randomised studies because treatment assignment may be determined by observable and unobservable factors. Since the outcome of interest is the result of the effect of the treatment after controlling for differences in the prognostic factors, a simple comparison of outcomes between groups may not provide an unbiased estimate of treatment effect when the prognostic factors are unevenly distributed between treatment groups. The consequences of selection and confounding bias are discussed in more detail in Section 2.1.

A variety of methods have been developed to minimise the risk of bias when making inferences on treatment effect using non-randomised studies. Most methods aim to make treatment and control groups comparable (matching methods, inverse probability weighting) or control for the effect of prognostic factors on outcome (regression adjustment, multivariate regression, propensity score, or instrumental variables). Some methods control for the effect of prognostic factors by using natural experiments that may be assumed to mimic randomisation (difference-in-differences and regression discontinuity). More sophisticated methods attempt to model the process of selection into treatment and the effect of the treatment on the outcome jointly (structural models, control function, correction approach). The most common methods to analyse comparative individual patient data are described in Sections 2.2 to 2.4.

The 2013 NICE Guide to the Methods of Technology Appraisal (Methods Guide)[2] recognises that non-randomised studies are at greater risk of bias and recommends that inferences should be made with care. However, the guide does not provide formal or detailed guidance regarding their use to inform estimates of treatment effect or the appropriate methods and

analyses to deal with any potential biases. At the same time, there is a growing interest in exploiting these studies to obtain estimates of treatment effect to inform cost-effectiveness analyses, either as a complement to RCTs or as sole source of data when no other evidence is available. Consequently, companies and evidence review groups (ERGs) may be unclear about the role of this evidence in different circumstances, how to choose between (and apply) the different methods available and how to critically review the methods used and the results of the analyses. Similarly, the appraisal committee may lack confidence in the robustness of estimates of treatment effect from non-randomised studies. Lack of clear guidance on the use of evidence from non-randomised studies may lead to its inappropriate use and the derivation of biased parameter estimates. This may have potential adverse consequences for decisions on the effectiveness and cost-effectiveness of technologies.

1.2. OBJECTIVES, SCOPE AND STRUCTURE

This Technical Support Document (TSD) summarises the methods available to analyse comparative individual patient data (IPD) from non-randomised controlled studies to obtain estimates of treatment effect to inform NICE Technology Appraisals (TAs). It proposes a set of recommendations to improve the quality and transparency of future assessments. The scope of this TSD are those methods that can be used to analyse IPD from subjects assigned to the intervention and a relevant comparator with the objective of obtaining estimates of treatment effect (i.e. the change in outcome attributable to treatment). The study designs that fall within this scope (i.e. those which produce comparative non-randomised data) are: quasi-randomised trials, quasi-experimental trials, controlled before-and-after, concurrent cohort, historical cohort, case-control and controlled cross-sectional. IPD from different sources on the intervention and the comparator can be used to form a comparative IPD dataset if the individuals can be assumed to be drawn from the same patient population and be exposed to similar confounders. Also within the scope is the critical appraisal of studies reporting estimates of treatment effect using non-randomised data.

There are a number of relevant topics to NICE TA outside the scope of this TSD:

- Studies with objectives other than estimating treatment effect, such as epidemiological studies aiming at establish the natural history of a condition or the link between intermediate and final outcomes
- IPD from single arm studies (either collected in a RCT or in a non-randomised study)

- Aggregate comparative data from non-randomised studies
- Appropriate study designs for inference of treatment effect.

These topics are potentially relevant to NICE TAs and therefore may be the focus of future TSDs. However, given the wide range of methodologies and the different features of these studies, it was decided to restrict the scope of this TSD to methods to analyse comparative IPD to obtain estimates of treatment effect.

This TSD is organised as follows. Section 2 provides a non-technical description of the most common methods for analysing comparative IPD from non-randomised studies to inform estimates of treatment effect. The objective of this TSD is to draw attention to the main (often overlooked) assumptions each method relies on. Section 3 reviews a sample of NICE TAs that have used non-randomised studies to obtain estimates of treatment effect, later used in the cost effectiveness analysis used in the submission. Section 4 reviews a selection of checklists to assist the critical appraisal of non-randomised studies and proposes a new checklist to address the gaps in the reviewed checklists and help assess whether the analysis of non-randomised data to estimate treatment effect is of adequate quality to inform decision-making. Section 5 summarises the key points for analysts and decision makers and suggests a number of areas for further research.

2. METHODS TO ESTIMATE TREATMENT EFFECTS USING NON-RANDOMISED DATA

2.1. BACKGROUND

The modern approach to the estimation of treatment effects is based on the counterfactual framework referred to as the *Rubin causal model* in recognition of his contribution.[3] Neyman[4] amongst others also proposed a similar statistical model. In the econometrics literature models involving counterfactuals have also been developed independently following Roy.[5]

There are two implicit assumptions in the treatment evaluation literature based on the Rubin causal model: (i) that the interest is on the evaluation of treatments which have been already experienced and (ii) that interest lies only in the mean causal effect. This contrasts with the more general econometric literature on policy evaluation,[6] which instead seeks to understand the "causes of the effects" in order to predict the impact of the intervention in other groups of people, time periods, etc. or even to predict the impact of new policies. A clear advantage of the treatment evaluation literature is that very few functional form restrictions are needed as it is only trying to identify a small number of causal parameters. The more general policy evaluation econometric literature can estimate a much larger set of parameters but it requires more structure for identification. *Identification* in this context refers to the ability to recover the causal parameter of interest from the estimated model. For example, if data are only available on the combined effect of two drugs given together, it will not be possible to separate (identify) the individual effects of each drug on the outcome. However, the individual effect of each drug may be estimated if some (structural) assumptions are made, such as that the two drugs have the same size effect. Any effects estimated rely heavily on these extra assumptions. Therefore, these additional assumptions must be justified with reference to other published studies, expert opinion, etc, and the sensitivity of the results to alternative assumptions should be examined.

2.1.1. The problem of selection

There is a fundamental problem when trying to evaluate a treatment. Each individual has two potential outcomes, one with and one without treatment. The *individual* treatment effect is the difference between these two outcomes. However, no individual is observed in both the treated and non-treated state at the same point in time and therefore the counterfactual is not

observed. One solution is to randomise the treatment between individuals. Although at the individual level the counterfactual is unobservable, at the aggregate (group) level randomisation of the treatment is intended to guarantee that any difference in mean outcomes between the treatment and control groups represents an appropriate estimator of the *average* treatment effect. It follows that the analyst can now infer that any difference in outcomes between the treatment and control groups can be ascribed to the treatment itself. However, appropriate randomisation is not always possible and in these cases a treatment or policy will need to be evaluated using non-randomised data.

Unlike data from RCTs, non-randomised data are generated in an uncontrolled environment. Non-random treatment assignment complicates the estimation of the treatment effect because the analyst cannot rule out the possibility that a patient received a particular treatment because of some (observable or unobservable) factors. This leads to the potential for selection bias in the estimation of treatment effect. Selection bias arises from differences in the characteristics that have an independent influence on the outcome between the individuals in the treated and the control groups.

Statistically, selection bias occurs when the treatment variable is correlated with the unobservables in the outcome equation. This correlation between the treatment variable and the unobservables can be the result of two issues. First, from incorrectly omitting observable variables that determine both the treatment and the outcome. This is referred to as 'selection on observables'. Second, from the presence of unobserved factors that determine both the treatment and the outcome. This is referred to as 'selection on unobservables'. The unobserved factors may be variables that were not collected in the dataset (and hence unobservable for the analyst) or variables that could not be measured. The correlation between the treatment and the unobservables is also referred to as 'endogeneity'.

The correlation between the treatment and the unobservables leads to inconsistent estimates of the parameter of interest. In other words, the estimated parameter measures only the association between the treatment and the outcome rather than the size and the direction of the effect (the causal relationship). For example, in his seminal paper McClellan *et al.*[7] set out to investigate if intensive treatment of acute myocardial infarction reduced mortality in the elderly using observational data. The risk of bias here lies in that patients may have lower mortality rates not as a result of the better treatment but due to their unobserved

characteristics, which in turn are related to the treatment received. If this is the case, the estimated treatment effect includes not only the effect of the treatment but also the effect of the unobserved characteristics on mortality. Whether the treatment appears more or less effective than it truly is depends on the type of model and the signs and relative sizes of the correlations between the outcome, the treatment and the unobservables.

It is important to emphasise that the term 'selection on observables' does not imply that the treatment assignment does not depend on unobservables, but rather that the unobservables which determine the treatment are not correlated with the outcome. Similarly, the term 'selection on unobservables' does not imply that treatment assignment does not depend on observable variables. Treatment assignment depends on both observable and unobservable characteristics but the critical difference is that those unobservable characteristics are correlated with the unobservables in the outcome equation.

2.1.2. Types of treatment effects

The treatment evaluation literature mainly focuses on a certain aspect (typically the mean) of the distribution of the individual treatment effects in the population (or a subgroup of the population) rather than the full distribution. There are several treatment effects which can be defined and might be of interest such as, the average treatment effect (ATE), the average treatment effect for the treated (ATT), the local average treatment effect (LATE) and the marginal treatment effect (MTE). The treatment effect which is typically of interest in NICE TAs is the ATE. The ATE measures the expected gain from the treatment for a randomly selected individual. In other words, the ATE calculates the expected effect of the treatment if individuals in the population under consideration were randomly allocated to treatment and is therefore the effect that would be identified by a RCT. This parameter is the most difficult to identify in general in the sense that it requires more demanding assumptions for identification than alternative treatment effects. These assumptions, as highlighted earlier, are often untestable and require thorough justification. An assumption that is easy to justify for a small group of similar individuals may be difficult to justify for the whole population. In this case the analyst could redefine the population of interest to the group of similar individuals but then the ATE obtained will not necessarily be valid for the initial population. In addition, the ATE is relevant where the treatment is applicable to the entire population represented by the data.

The ATT is relevant when the interest lies on the effect of the treatment only for those who are treated. This could be the case if, for example, it is known that certain individuals are unlikely to benefit from the treatment. Therefore they are almost never treated and their treatment effect is of no interest.

The LATE[8] is 'local' in the sense that it measures the ATE on a particular subpopulation of similar individuals but does not apply to the whole population. This treatment effect is easier understood in the context of instrumental variable methods. Section 2.3.1 returns to this concept and provides a more thorough explanation.

Treatment effects can be homogeneous or heterogeneous. A treatment effect is homogeneous if the treatment has the same effect on individuals who differ in both observed and unobserved characteristics. It follows in this case that the ATE, ATT and LATE are identical. However, individuals may vary in their responses to a treatment. This is referred to as heterogeneous treatment effects or response heterogeneity. This heterogeneity may be determined by unobserved and observed characteristics. In most cases, selection into treatment will depend on the determinants of this heterogeneity and this may lead to differences between the treatment parameters, ATE, ATT and LATE.

The MTE[9] is another treatment effect that is used in the literature. It is defined as the average treatment effect for those individuals who are indifferent between receiving the treatment or not at a given value of the observable variables. The MTE can be used to construct all the treatment effects defined above and it is especially useful when individual heterogeneity in the treatment effect is important for the decision problem. Methods dealing with individual heterogeneity in the treatment effects are outside the scope of this TSD. More details of these methods and their usefulness in the evaluation of treatment effects can be found in Basu *et al.*,[10] Basu[11] and references therein.

Different estimators place different restrictions on the counterfactuals that can be identified and hence on the treatment effects that can be consistently estimated. It is important to define the treatment effect of interest before attempting estimation but it is also important to recognise that, in some cases, the assumptions required to identify the parameter of interest cannot be justified and only an alternative treatment effect can be estimated. In other words,

there may be situations where the treatment effect of interest is ATE but only ATT or LATE can be estimated.

The following sections give a non-technical description of the most commonly used methods to deal with the problems presented by non-randomised data in the literature. They highlight the major assumptions in each of the methods and their similarities and differences. There are excellent technical journal articles and reviews of the methods (see for example Blundell and Costa-Dias, Imbens and Wooldridge, Wooldridge, Jones and Rice,[12-15] and references therein; Nichols[16,17] gives some practical guidance on checking assumptions using STATA). The literature in this area is however in constant development and more advanced methods are continuously emerging.

Section 2.2 provides an overview of general issues which are important when estimating treatment effects. Section 2.3 concentrates on the methods that assume selection on observables. Section 2.4 describes methods that can be used when the assumption of selection on observables is thought to be untenable and there is good reason to believe that there are some unobservables which affect both the treatment and the outcome. Section 2.5 presents methods that can be used when data originated from natural experiments or quasi-experiments.

2.2. GENERAL CONSIDERATIONS

This section provides a summary of issues that must be taken into account when trying to estimate a treatment effect. Some of these issues are specific to this literature, others are applicable more generally to any statistical model building study.

First, the analyst must have a clear understanding of the process by which individuals are assigned or take up the treatment to be able to select the appropriate method to estimate the treatment effect, given that alternative methods use different assumptions. These assumptions should be clearly discussed and justified since the estimated treatment effect depends heavily on those identifying assumptions. Assumptions that might be perfectly reasonable in one case can be indefensible in a different scenario. In some cases, placebo tests can be used as a robustness check on the identification strategy as outlined by Jones.[18] For example, Galiani *et al.*[19] set out to assess the impact of privatisation of water services on child mortality in Argentina. They found a reduction in deaths from infection and parasitic diseases and

perinatal deaths in those municipalities that had privatised their services. As a check, they used the same model to find the impact of the privatisation on deaths from causes known to be unrelated to water conditions such as accidents, cardiovascular diseases, etc. and they found no effect. Those results gave support to their identification strategy. Given that these assumptions are in general untestable, the impact of different assumptions should be tested in the sensitivity analysis. It is also good practice to compare the results obtained to those in other studies with similar datasets. Questions 1 to 3 of the QuEENS checklist in Section 4.2.2 highlight these general issues. The structural uncertainty arising from the different plausible models can be accounted for in decision models by model averaging as described in Jackson *et al*.[20]

In addition, if a parametric model is used, all the issues relating to general good model specification and testing also apply.[21] These are over and above those checks specific to ensuring that the treatment effect is properly specified. For example, if a parametric model is used for the outcome model, it is important that the model is consistent with the outcome variable. Linear regression might be appropriate for unbounded outcomes, for binary outcomes logit or probit models might be good candidates, generalised linear models can be useful in cases where the data is highly skewed, etc. Some examples of models which take into account the problems raised by the data typically used in or to inform economic evaluations can be found in Nixon and Thompson,[22] Basu and Manca,[23] Hernández *et al.*,[24] Jones *et al*.[25] As with any other statistical modelling study, the analyst should look carefully at the estimated models as issues such as counterintuitive results or parameters might be a consequence of model misspecification. Questions 4 and 5 of the QuEENS checklist in Section 4.2.2 highlight these general issues.

Finally, access to good quality data is essential for all approaches to obtain reliable estimates of the treatment effect.

2.3. METHODS ASSUMING SELECTION ON OBSERVABLES

The methods described in this section rely on the *ignorability of treatment* (conditional on a set of observed variables), also known as *unconfoundedness* or *conditional independence*.
That is, once conditioned on (controlled for) a properly chosen set of observed covariates, the potential outcomes are independent of the treatment assignment, as is the case when the treatment is randomised across individuals. A weaker version of this assumption requires

only *ignorability in mean.* That is, the mean of the distribution of potential outcomes without treatment is the same regardless of treatment status.[i] The same applies to the mean of the distribution of potential outcomes with treatment. Intuitively, the potential outcomes can be assumed to be mean independent of the treatment if there are enough observed variables which determine treatment selection and those observed variables are controlled for. This weaker version of ignorability (together with additional assumptions) is sufficient for the identification of the ATT.

These methods require a good understanding of the process of selection into treatment and this is largely dataset dependent. It is therefore crucial in any empirical application to support this assumption using, for example, published literature and/or expert opinion. The methods work best when there is a rich set of relevant covariates which also adds plausibility to the ignorability assumption. Care should be taken not to include variables in the conditioning set which can be affected by the treatment as it would in general cause ignorability to fail. Suitable variables to include are variables measured before assignment to treatment which might even include past values of the outcome of interests. However, variables which predict treatment but are independent of the unobservables that affect outcomes (i.e. *instrumental variables*) should not be included. Inclusion of instrumental variables will increase the bias in the treatment effect estimate, unless they are exogenous. Even in this case, when they do not cause a bias, they will increase the variance of the estimate and will thus be inefficient.[26] This important assumption cannot be tested. Therefore, it should be justified with reference to the expert literature and opinion. In some cases, it can be tested indirectly (see Imbens and Wooldridge's[13] for details).

The *overlap assumption* is required in addition to the *ignorability assumption* in order to identify ATE. Overlap means that, for any combination of covariates, there is always the chance of seeing individuals in both the treatment and the control groups. It rules out the possibility that some individuals with certain observable characteristics are always in one group and never in the other. It is not possible to estimate the ATE for a dataset that includes individuals with those characteristics.

i Mathematically this can be stated as $E(y0|X,d)= E(y0|X)$ where E is the mathematical expectation, y0 is the outcome without treatment, X is the matrix of control variables and d is a binary variable indicating treatment assignment

Rosenbaum and Rubin[27] refer to ignorability plus overlap as *strong ignorability*. This assumption is fundamental when trying to identify the ATE. Weaker versions of both assumptions are enough to estimate the ATT. These are the ignorability in mean for the potential outcome without treatment and a version of the overlap assumption which only requires that the treated individuals have a chance of also being in the control group. Assuming ignorability holds, lack of overlap in the covariate distribution is an important issue which needs to be assessed thoroughly and it is probably the biggest issue that the analyst will face. Matching as well as trimming, that is, selecting a population with overlap,[28] may sometimes be used to deal with problems of overlap. Details of possible checks to use to assess this important assumption can be found in Section 4.2.2 Question 7.

2.3.1. Regression adjustment

Regression adjustment (RA) uses a regression model conditional on covariates to predict the outcomes. It fits two separate regression models for the treated and untreated samples and the treatment effects are then based on the difference between the predictions of the two models. The regression model can be very flexible as the means are non-parametrically identified. Therefore, in addition to the usual parametric regression models (linear, probit/logit, Poisson, generalised linear models, etc.), nonparametric estimators such as kernel or series estimators can also be used. Note that this implies a two step procedure. First, two regression equations are estimated, one for the treated group and another one for the control group. In a second step, the individual differences in the predictions for the two potential outcomes are averaged across all individuals. In this case, the analyst needs to appropriately correct the standard errors in the second step to take into account that the potential outcomes in the first step are estimated. This can be handled by using any of the standard methods for two steps estimators. Alternatively, the analyst can estimate both steps jointly (see for example the command "teffects ra"[29] in STATA 13).

One problem that sometimes is ignored when using parametric models in this setting is the lack of common support (or overlap) in the set of covariates used in the regression model. As discussed earlier, overlap, also known as matching assumption, is required to accurately estimate the treatment effect when predicting the average outcome from the two regression equations. Essentially, it ensures that the treated and control cohorts are similar. However, as parametric methods can be estimated despite having overlap problems, sometimes analysts overlook checking this important assumption. One of the advantages of using nonparametric

estimators is that the analyst is required to tackle any problems with overlap in the covariate distribution as part of the estimation of the nonparametric model.

Regression analysis or covariate adjustment is a simpler form of RA which restricts the parameters and variances of the regressions for the treated and untreated to be the same only allowing the constants to differ. Although in many applications the treatment effect is assumed to be homogeneous, treatment-by-covariate interactions can be introduced to allow for some forms of heterogeneity in the treatment effect. Interactions assume that the heterogeneity in the treatment effect is dependent on observable covariates but there is no unobserved heterogeneity.

If a parametric model is used, the mean function chosen should be consistent with the outcome variable. For example, a linear regression might be appropriate for unbounded outcomes but not for binary outcomes. For binary outcomes, logit or probit models are good candidates. As with any statistical analysis, the model should be checked for misspecification in the model.

2.3.2. Inverse probability weighting

Inverse probability weighting (IPW) is the first method presented here that uses the propensity score function. The propensity score is the probability of treatment assignment as a function of a set of observable covariates. The ATE corresponds to the difference in the weighted means. These weighted means are calculated using the inverse of the propensity score as weights.

The IPW is an estimator of the treatment effect that corrects for missing data. Given that we have selection on observables, some observations will be very likely to appear in the treatment group but very unlikely to appear in the control group (i.e. they are likely "missing" from the sample) and vice versa. The IPW estimator compensates for this by giving more weight to the small number of observations which appear on one group but have a small probability of being found in that group. It uses the inverse of the propensity score to calculate weighted means for the treated and control samples. In this way, observations that are similar to those which are likely to be missing receive a higher weight in the computation of the sample mean. The usefulness of this method depends on how well the model for the propensity score predicts the probability of treatment. In other words, that the propensity

score is properly specified according to the usual specification test and therefore, there are no systematic problems with the predictions of the probability of treatment.

The propensity score function should be sufficiently flexible.[27] Flexibility is required to ensure the model is able to produce non-linear associations. Flexible forms for the propensity score will tend to rule out cases where the propensity score is zero or one. This signals a failure of the overlap assumption. Nonetheless, overlap should still be checked for independently. The flexibility can be achieved by using a parametric model such as a probit or logit and including functions of the covariates such as squares, polynomials and interactions.

Two important issues arise if there are problems with overlap. It is possible that some of the predicted propensity scores are close to zero, implying an excessively large weight rendering the IPW method unstable (see Radice *et al.*[30] and Kreif *et al.*[31] for the potential consequences of unstable weights). At the same time, the usual parametric models for binary data, the probit and logit models, which tend to give very similar predictions in the middle of the propensity score, are more likely to differ when the propensity scores are close to zero or one. The choice of parametric model, therefore, might become important and have an impact on the results.

2.3.3. Doubly robust methods

RA estimates a model for the outcome but it does not model treatment selection. On the other hand, IPW estimates a model for the probability of receiving treatment but it does not model the outcome. Doubly robust methods combine RA and IPW. The advantage is that only one of the two models needs to be specified correctly to be able to identify properly the treatment effect. However, if both models are misspecified, they have been shown to produce biased estimates of the treatment effect.[32]

There are two possible estimators which combine the RA and the IPW methods. The augmented IPW is an IPW estimator with an augmentation term to correct the estimator for misspecification in the treatment model. Similar to the IPW model, this estimator will be unstable when the predicted treatment probabilities are close to zero. An alternative estimator is the RA-IPW. This estimator is based on the RA estimator but it makes use of the inverse probability weights to correct the estimator for misspecification of the regression function.

Alternative doubly robust estimators, uses and comparisons can be found for example in Kang and Schafer,[33] Robins *et al.*,[34] Kreif *et al.*,[35] Petersen *et al.*[36]

2.3.4. Regression on the propensity score

Regression on the propensity score assumes that the propensity score is enough to control for the correlation between the treatment and the covariates. The simplest method uses a parametric regression for the outcome variable but kernel and series estimators can also be used. Consistency of the ordinary least squares estimators relies on correctly specifying the model for the propensity score. Regression on the propensity score is, in general, inefficient.[14] Examples of regression on the propensity score in cost-effectiveness analysis can be found in Mitra and Indurkhya[37] and Manca and Austin.[38]

2.3.5. Matching

Matching aims to replicate randomisation by identifying *control* individuals which are similar to the *treated* in one or more characteristics. Matching estimators do not require parametric assumptions for the treatment effect. Matching identifies the treatment effect by using data from a group of comparison individuals with observable characteristics matching closely those of the treated individuals. That is, it compares the outcomes of individuals who differ in the treatment variable but are otherwise observationally very similar. The individuals used for the comparison may even come from a different population to the treated individuals as long as the assumption of selection on observables is still plausible. Matching methods identify the ATT parameter under fairly weak assumptions but need stronger assumptions to identify the ATE parameter in common with the rest of the methods in this section (see general introduction to Section 2.2). Matching works best if there are a large number of individuals to use in the matching cohort, a large number of covariates to model the propensity score and when the treated and control groups come from the same environment (bailiwick).

Matching can be exact or inexact. Exact matching is only feasible for a small number of discrete covariates and a large sample size. Therefore, it limits the validity of the ignorability assumption. Given that it is only feasible to use a small number of covariates, matched individuals might not be as similar as they need to be. Inexact matching generally employs a scalar (i.e. a number) obtained as a function of the covariates to match individuals. The two most popular inexact matching methods are nearest neighbour matching and propensity score matching. Propensity score matching uses the propensity score as the basis for matching

individuals. Nearest neighbour matching typically uses a multivariate measure of distance (typically the Mahalanobis distance[14]) to identify matches that are as close as possible to the treated individual on the basis of each covariate determining treatment assignment, but can also be defined in terms of the propensity score to reduce the dimensionality problem of having to match on a high number of covariates. Matching on the propensity score requires adjustment of the treatment parameters because the propensity score is estimated rather than observed. Abadie and Imbens[39] derive the adjusted standard errors which are also implemented in STATA 13.[40]

There are a number of decisions required to implement matching estimators. Analysts should record and justify their choices. For example, the implementation of matching on propensity score requires decisions on whether to match with or without replacement, how many individuals to use in constructing the counterfactual and what matching method to use. Matching without replacement means that each individual in one group will be matched at the most to one individual in the other group. Matching with replacement allows multiple matches so that, for example, the same individual in the control group can be matched to more than one treated individual. Matching without replacement might lead to poor matches in terms of the propensity score if one group is small compared to the other group.

Identifying the ATE requires ignorability and overlap so that the probability of being treated for any individual is never zero or one. Identifying the ATT requires ignorability in mean but only for the outcome without treatment and a weaker version of overlap where the probability of treatment might be zero for some individuals. Consequently, one needs to construct/impute the counterfactuals for both the treated and the control individuals when computing the ATE but only the counterfactual for the treated individuals when computing the ATT. Therefore, dropping control and/or treated individuals will make it more difficult to interpret the estimated parameter as an ATE. In contrast, for the estimation of the ATT, dropping treated individuals changes the population covered by the estimated treatment effect (however, dropping controls does not). Matching with replacement overcomes these problems but might lead to the same individual in the control group being used a large number of times in areas where the number of controls is small. The trade-off between the bias and the variance of the matching estimator should also considered when deciding the number of matches. Using only the closest match ensures that the bias is reduced. However, the variance of the estimator decreases by using more matches in the control group. At the same time, the bias increases if

the additional matches are not as close as the first. A compromise solution often used in practice is to only use those matches within a pre-specified radius (caliper matching).[41]

Other alternative matching estimators are kernel matching, interval matching and more recently genetic matching.[42] Kernel matching employs a kernel weighted average of individuals to construct the counterfactual and can also be defined in terms of the propensity score. Propensity scores are also used in stratification or interval matching where the range of the propensity score is divided in intervals. The intervals are calculated so that both the treated individuals and the controls have on average the same propensity score. Genetic matching uses a genetic search algorithm to find a set of weights for each covariate such that optimal balance is achieved after matching. Both propensity score matching and matching on the Mahalanobis distance have been shown to be special cases of this more general method.[42]

Lack of overlap will cause problems for matching estimators. Therefore, the overlap assumption should be assessed before and after matching (see Question 7 of the QuEENS checklist in Section 4.2.2). Trimming the sample to a region with overlap is an option for when overlap is still inadequate after matching. However, trimming may redefine the patient population on which the effect of the treatment is being estimated. As a result, it might no longer be considered generalisable to the original patient population defined pre-matching.

Stuart[43] provides a good overview of matching methods. Practical guidance on propensity score matching can be found in Caliendo and Kopeinig.[44] Examples of matching in clinical and cost-effectiveness evaluations can be found in Manca and Austin,[38] Austin,[45] Sekhon and Grieve,[46] Kreif *et al.*,[31] Radice *et al.*[30]

2.3.6. Parametric regression on a matched sample

Matching estimators may also be combined with regression adjustment. Parametric regression on a matched sample controls for any remaining imbalances between the treatment and control group after matching. Kreif *et al.*[32] show that regression on a matched sample report less bias than doubly robust methods in situations where the IPW is unstable and both the outcome and the propensity score models are misspecified.

2.4. METHODS FOR SELECTION ON UNOBSERVABLES

None of the methods described in the last section can identify the treatment effect if there is selection on unobservables, also referred to as endogeneity of the treatment variable. Perhaps the most common method to deal with endogeneity in the treatment evaluation literature is the *Instrumental Variable* (IV) approach described below. More complex models are also available such as *structural models* in econometrics but require a higher level of expertise and often bespoke programming.[14]

2.4.1. Instrumental Variable methods

IV methods can prove very valuable in estimating the treatment effect when the unconfoundedness assumption is suspect. The strategy in IV estimation is to find a variable (referred to as an 'instrument') which is correlated with the treatment but only correlated with the outcome through its effect on the treatment. This is known as the exclusion restriction, as the IV is exclusive to the treatment choice equation. The variation in the instrument is then exploited to identify the causal treatment effect. Intuitively, since the instrument is a source of variation correlated with the treatment decision, it gives some exogenous variation in which to approximate randomisation. It follows that the choice of an appropriate instrument is fundamental. McClellan *et al.*[7] set out to investigate if intensive treatment of acute myocardial infarction reduced mortality in the elderly using observational data. They used measures of the distance between residence and hospital as instrumental variable. They hypothesised that distance between residence and hospital was correlated with how intensively a patient would be treated for acute myocardial infarction but mortality was not directly associated with this distance. Patients with similar observable characteristics who only differ in the treatment intensity due to factors unrelated to the outcome (these factors being potential instruments), such as distance to hospital, can then be used to estimate the treatment effect. Often, in practice, it is difficult to find an instrument that satisfies the exclusion restriction. In addition, the instruments might not have sufficient variation themselves or induce enough variation in the treatment in practice. These are known as weak instruments. In this case, the model is said to be weakly identified, leading to a loss of precision in the estimates of the treatment effect. It is extremely important to ensure that an instrument is exogenous if it is weak. An exogenous variable is one which is determined outside the model, that is, it is external to it. In the example above, the distance between the place of residence and the hospital is exogenous in the model which explains mortality. Even moderate levels of endogeneity of the instrument can lead to estimate parameters that are

more inconsistent than those obtained from methods which wrongly assume selection on observables. Recent work[47] proposes the use of a diagnostic criterion to determine the circumstances under which the IV estimator would produce a better estimate of the true causal effect than an ordinary linear regression (OLS) using prior knowledge.

A key assumption in standard IV estimation is that the treatment effect is homogeneous (the treatment effects are the same for everybody in the population). Alternatively, the treatment effect can be assumed heterogeneous but requires the assumption that selection into treatment is not influenced by the unobserved heterogeneity (individual specific effect) in the outcome. Only with these assumptions is the IV estimator able to recover the ATE. If the unobserved individual effect in the outcome is correlated with treatment receipt, traditional IV methods can identify the local ATE (i.e. LATE). This is valid as long as the probability of treatment is a monotonic function of the instrumental variable. That is, the probability of treatment never increases or never decreases with the instrument. The estimated LATE is local in the sense that it measures the effect of the treatment on those who are induced to take up the treatment by the change in the instrument. The monotonicity assumption implies that the instrument induces a change in treatment in the same direction for all individuals involved. Clearly, the LATE depends on the instrument. Therefore, different instruments will give different values of the LATE if they induce different groups of people to engage in treatment.

Local IV methods have been proposed in the literature to overcome the limitations of the traditional IV approach (see Basu *et al.*[10] for an application in health economics). These methods aim to estimate the whole distribution of treatment effects using the MTE. Once the distribution is identified, the analyst can calculate all the aggregate parameters such as the ATE as weighted averages of MTE.

2.4.2. Panel data models

Panel data offers the advantage of using the individual as their own control since each individual is observed at different time periods. The assumption required to be able to identify a treatment effect is that the individual unobserved heterogeneity in the outcome equation is time invariant. If the individual unobserved effect is suspected to be correlated with the covariates in the model (including treatment), a fixed effects model or a first difference model can be used (see next section for the relationship between the first difference model and the difference-in-difference model with longitudinal data). These

models do not use the cross-sectional variation across individuals and might therefore be less efficient than the random effects estimator. This estimator however, assumes that the unobserved determinants of heterogeneity in the outcomes do not have an effect on selection. The Hausman test[48] of fixed versus random effects is routinely reported by standard software when using these models. It is worth noting that rejection of the null hypothesis of this test could also be due to misspecification of the model. With the exception of the models described above, estimation of treatment effects using panel data tend to be complicated by dynamics and are outside the scope here (see Wooldridge,[14] Jones and Rice[15] for a brief account of additional considerations and further references).

2.5. NATURAL EXPERIMENT APPROACHES

Natural experiment approaches make use of exogenous events. Exogenous events are those that induce a random assignment of individuals to treatment or to eligibility for treatment. That is, purely by chance, individuals end up in the treated and control groups. Natural experiments have been used to evaluate population health interventions and guidance on their use has been published by the Medical Research Council.[49] As pointed out earlier, natural experiments can generate instruments but there are other approaches that can also be used to estimate a treatment effect in these cases.

2.5.1. Difference in differences

The usual difference-in-differences (DID) approach compares the difference between the treatment and control groups before and after the natural event. To identify a treatment effect, DID uses either longitudinal data for the same individuals or repeated cross-sections drawn from the same population, before and after the treatment. By comparing the changes over time in the means of the treatment and control groups, the DID estimator allows for both group-specific and time-specific effects. In general, the DID approach will identify the ATT. Bellou and Bhatt[50] studied the effect of changing the design of identification cards for under 21s in the US on alcohol and tobacco use. Between 1994 and 2009 forty-three US states changed the design of their driver's license/state identification cards with the purpose of reducing underage consumption of alcohol and tobacco. The study was able to use a DID methodology by exploiting the fact that different states introduced the change in design at different points in time creating treatment and control groups. A significant reduction in consumption was found but only in the first couple of years after the introduction of the design change.

The DID estimator makes two important assumptions. First, it assumes common trends across the treatment and control groups. If the treatment and control groups come from very different samples which might be subject to different shocks, the assumption of common trends may be questionable. The possibility of failure of this assumption motivates the differential trend adjusted DID estimator[51] which needs historical data with a similar trend before treatment. Second, although it allows for selection on unobservables, it restricts its source by assuming no selection on the unobserved time varying individual specific shocks. This assumption will fail if, for example, individuals behave differently in anticipation of the treatment, causing a fall in the outcome variable (the "Ashenfelter's dip"[52]). In addition to these two assumptions, when using repeated cross-sections of individuals instead of longitudinal data, there is an implicit assumption of no systematic changes in the composition of the groups so that the average individual fixed effect can be eliminated.

When longitudinal data are available, the DID estimator is the first difference estimator used in panel data to obtain consistent estimates in the presence of fixed effects, that is, individual specific, time-invariant effects which are correlated with the covariates in the model.

The DID approach has been extended to nonlinear models when, for example, the outcome is a binary variable. The DID estimator can also be combined with matching to help relax some of the assumptions of both methods.[53]

2.5.2. Regression discontinuity design

The regression discontinuity (RD) design exploits discontinuities in treatment assignment due to a continuous variable (the forcing variable). For example, eligibility might be function of age. The RD estimator uses the discontinuity to identify a treatment effect by assuming that the individuals on different sides of the discontinuity are the same in terms of the unobservables that affect the outcome, and that treatment differs simply because of the discontinuity in the eligibility rule. The discontinuity may be 'sharp' if treatment assignment is a deterministic function of the forcing variable. For example, if all individuals above a certain age threshold are treated but those below the threshold are not. Alternatively, the discontinuity may be 'fuzzy' if the probability of treatment is discontinuous in the forcing variable but does not determine perfectly treatment assignment. For example, if other variables (including unobserved variables) apart from age determine treatment. In this case, treated as well as untreated individuals are found at both sides of the threshold. Both designs

rely on the assumption that the conditional mean of the outcome functions for the treated and control groups are continuous at the discontinuity point.

Zhao *et al.*[54] use a regression discontinuity approach to assess the impact of giving individuals information about their actual health status on their diet. They took advantage of a Chinese longitudinal dataset which included data collected from a physical examination of every individual in the survey. Specifically, trained examiners measured the blood pressure of every individual who was then informed of the results. Since all the individuals were informed (the treatment), there is no self-selection into treatment. In addition, hypertension is a deterministic function of blood pressure measures, that is, either the systolic blood pressure and/or the diastolic blood pressure are above certain limits. Given these two features a regression discontinuity design approach can be used to assess the causal effect of information on diet. The study found a significant reduction of fat intake after receiving a hypertension diagnosis.

Note that the sharp design implies that treatment is exogenously determined by the forcing variable. Therefore, that, at least in the neighbourhood of the discontinuity, there is no selection on unobservables affecting the outcome. However, matching cannot be used because of the lack of overlap between the treatment and control groups. Other methods, such as regression adjustment, will rely on dangerous extrapolation due to the complete absence of overlap in the key variable(s) determining the discontinuity. The fuzzy RD design requires the additional assumption that there is no selection on individual specific gains at the local level to be able to identify a local treatment effect. That is, individuals below and above the threshold are similar in terms of unobservables. This is however a strong assumption even at a local level.

The parameter identified by the RD estimator is a local average treatment effect similar to the LATE parameter identified by the IV estimator described above. In a sharp RD case, the estimated parameter coincides with that estimated using IV methods. The parameter identified in the fuzzy RD case and by the IV estimator are the same if they are applied to the same neighbourhood of the discontinuity. However, their interpretation is different. If it is likely that individuals have no influence in treatment selection at the local level, then the estimated treatment effect represents local effects on randomly selected individuals (the RD design interpretation). Otherwise, it represents the local treatment effect on those individuals

who are induced to take up the treatment (the IV LATE interpretation). In common with the IV method, a large sample size is needed to ensure precision of the estimated treatment parameters given that we are restricting the analysis to a local area.

2.6. SUMMARY

A key recommendation is that the analyst should have a clear understanding of the process of treatment assignment. This understanding should inform the method selection. The assumptions underlying the method chosen should be tested as possible and justified. In addition, the sensitivity of the results should be explored by estimating alternative models which rely on different assumptions.

Any method using a parametric model for either the outcome, the treatment (propensity score) or both should ensure that the chosen model is consistent with the dependent variable. Appropriate specification tests for each model should be carried out just as with any other statistical modelling exercise.[21]

Methods which assume selection on observables rely on good overlap in the distribution of the covariates. Lack of overlap in parametric models is often overlooked but can have serious consequences as the models extrapolate to regions outside the sample. It also causes instability when using IPW. An advantage of nonparametric regression models as well as matching is that it forces the analyst to consider overlap directly. Note that misspecification of the regression function might be even more problematic when the overlap is poor as misspecification of the parametric models tends to be more pronounced in areas where the data is scarce. If reasonable balance cannot be achieved then one solution might be to trim the sample to the region of common support. However, it is important to note that this redefines the group of individuals for which the treatment effect is estimated.

Methods which assume selection on unobservables rely on finding good and reliable sources of exogenous variation. The analyst needs to highlight the assumptions underpinning the analyses, testing them where possible.

3. REVIEW OF TECHNOLOGY APPRAISALS USING NON-RANDOMISED DATA

A pragmatic review of NICE TAs was conducted to understand how non-randomised data has been used to inform estimates of treatment effect and to motivate the subsequent investigation of alternative methods. TAs using non-randomised data to inform estimates of treatment effect were identified through an informal survey of Chairs of NICE appraisal committees, NICE technical leads and the members of the NICE Decision Support Unit (DSU). This was supplemented by an unpublished review of the use of non-randomised evidence in TAs in a selection of 110 TAs (from TA151). The final appraisal document (FAD) was used to ascertain whether and how non-randomised data was used to inform estimates of the treatment effect and this was complemented with information from the evaluation report (when available on the NICE website). The use of non-randomised data for other purposes, such as to obtain estimates of health-related quality of life or costs for health states or events, to inform the link between intermediate and final outcomes was not considered in this review. A data extraction form was designed and information was extracted on the parameter informed by the effectiveness estimate obtained from non-randomised data, the method employed and details on the methodology.

Appendix 1 summarises the results of this consultation. The review identified 16 TAs using non-randomised data to inform estimates of treatment effect. Five TAs used IPD on the interventions and comparators. TA130 and TA195 used non-randomised data from the British Society of Rheumatology Biological's Register to inform the transition between health states and the health utility gain from treatment. TA185 used IPD on the intervention from a Phase II trial and IPD on the comparator from other four Phase II studies (historical controls). TA279 used studies using US Medicare claims data to obtain estimates of the relative risk for death. TA304 used IPD from the National Joint Registry to estimate revision rate of prostheses.

Four TAs used IPD on the intervention. In TA188, the improvement in height associated with human growth hormone was informed by an analysis of the Kabi International Growth Database; however, no details on the methods employed were given in the assessment group report. TA202 and TA299 used IPD from a single arm Phase I/II study to obtain the hazard

ratio for overall survival and progression free survival in cancer patients. ID667 used compared the intervention arm of an RCT with aggregate data from the comparator to obtain the hazard ratio for survival.

Seven TAs used aggregate data from non-randomised studies. TA156 is an example of the use of estimates of treatment effect from published studies using non-randomised data. In TA165, the relative risks for primary non-function, delayed graft function, graft survival and overall survival were calculated by the assessment group based on the rates of events reported in published non-randomised studies. In TA166, the treatment effect was modelled as the improvement in health-related quality of life associated with cochlear implantation obtained from non-randomised studies. In TA209 and in TA241, the survival under the intervention and the comparator were obtained from individual arms of RCTs and cohort studies. In TA242, one of the companies' submissions presented a mixed treatment comparison that included randomised and non-randomised data to inform the survival benefits of an intervention. In TA246, the risk of systematic reaction following a sting with the intervention is the pooled risk observed in the patient groups using the intervention in the RCT and non-RCTs identified in a systematic review. The risk of systematic reaction without the intervention was obtained from a survey study.

Of the 16 TAs, six used multivariate regression to adjust for differences in characteristics between the patients on the interventions and the comparators (TA130, TA195, TA185, TA279, TA202, ID667) and two used propensity score matching (TA279, TA304). In seven TAs, no methods appear to have been used to adjust for potential differences. Of these, five applied unadjusted estimates directly in the model (TA188, TA299, TA165, TA209, TA246). Two TAs used estimates from non-randomised studies to inform evidence synthesis, which was then used to inform the treatment effect in the model (TA156, TA242).

The review of NICE TAs using non-randomised data to obtain estimates of treatment effect illustrated the types of data used: comparative IPD on the different interventions and comparators, IPD only for the intervention, and as aggregate data, either as an estimate of treatment effect obtained from a published study or aggregate estimates of outcome for the interventions or comparators. The TAs that used comparative IPD used multivariate regression and propensity score matching to adjust for (observed) differences between the intervention groups. Most of the TAs identified by the review used aggregate data from

published studies and obtained estimates of treatment effect unadjusted for potential confounders. This highlights the need for guidance in the analysis, interpretation and critical appraisal on the use of non-randomised evidence for TA.

This review has some important limitations. The identification of TAs had a pragmatic design. Reviewing all the TAs published so far was impractical within the time and resources available. For this reason, there may be other TAs that have used non-randomised data to inform estimates of treatment effect that are not included in the review. Secondly, the data extraction effort was concentrated on the FAD and complemented with the assessment report (where available on the NICE website). Therefore, there may have been details on the methods that have been missed from the review. Furthermore, the time constraints and the lack of detail in some TAs meant that it was not possible to make judgements on the appropriateness and quality of the analysis. Despite these limitations, the review has met its objectives of illustrating the use of non-randomised data to inform estimates of treatment effect and motivate the subsequent sections of this TSD.

4. RECOMMENDATIONS

The review of methods in Section 2 and the review of NICE TAs in Section 3 inform our recommendations for future NICE TAs:

- An algorithm to help inform the selection of the appropriate methods for the analysis of comparative IPD (Section 4.1).
- A review of the currently available checklists to help critically appraise published studies using non-randomised data (Section 4.2.1).
- A new checklist to help critically appraise the analysis of non-randomised data that addresses the gaps of the checklists previously identified (Section 4.2.2).

4.1. ALGORITHM FOR METHOD SELECTION

Figure 1 classifies non-randomised data according to whether it is comparative and whether it is available as IPD or in aggregate from a published study. Comparative non-randomised data can be available as IPD on the treated and control individuals or as estimates of treatment effect reported in a published study. The algorithm in Figure 2 and Figure 3 suggests a number of sequential steps to help choose the appropriate method to estimate treatment effect from comparative IPD. The quality of estimates of treatment effect reported in a published study will be discussed in Section 4.2. Non-comparative data, either IPD on treated or control or aggregate estimates on both from different non-randomised studies, are outside the scope of this TSD and hence not covered in the subsequent sections.

Our proposed algorithm for methods is depicted in Figures 2 and 3. Figure 2 summarises the methodological options available for non-randomised comparative IPD. Since it is almost impossible to know which method is best, results obtained from alternative methods that make alternative plausible assumptions should be presented as sensitivity analysis. Although the choice of method should be driven by the treatment effect of interest (ATE, ATT, other), the availability of data and the mechanism by which people were assigned to treatment or otherwise will play a central role in what can and cannot be identified. Hence, the first question to ask relates to the type of data the analyst has access to, namely whether IPD is from a natural experiment or quasi-experiment. IPD from natural experiments lends itself to relatively simple methods such as DID, regression discontinuity or even IV estimation if the natural experiment provides a valid instrument. As long as the assumptions behind these

models can be seen as reasonable and the estimated treatment effect is relevant, this would be the preferred approach to take in the first instance.

Often the IPD is unrelated to a natural experiment. In this case, the analyst should be able to explain the mechanism by which individuals are assigned to treatment and justify this with convincing arguments. Consequently, a key question to ask is whether it is reasonable to assume that the selection of individuals to the intervention or control (treated or untreated individuals) is related only to the observed variables. For example, in a study comparing surgery with medical management, the selection to surgery may be assumed to be related only to the clinical characteristics of the patients that are recorded in the dataset. This assumption cannot be formally tested. Its plausibility can be discussed in light of the existing evidence around the intervention, the clinical pathway, patients' and clinicians' preferences.

If the assumption of selection on observables is untenable, there are a few options depending on the data available and the assumptions that the analyst can reasonably make. Access to longitudinal IPD in which the same individuals are followed over a period of time opens the avenue of panel data models. If the parameter of interest is the treatment effect at a particular time period, standard models under the assumptions of individual time invariant heterogeneity and ignorability conditional on unobserved individual heterogeneity can be used. More sophisticated models are needed to relax the ignorability assumption.

Another possible option if the assumption of selection on observables is unlikely to hold is IV analysis. This type of analysis requires the availability of a variable (also known as instrument) that is correlated with the treatment assignment but unrelated to the outcome directly.

Figure 3 focuses on the methods that assume selection on observables: multivariate regression, regression adjustment, matching, IPW, propensity score matching and regression on propensity score. All these methods rely on a good overlap in the covariate distribution of the treatment and control groups. That is, for any combination of observable characteristics, there is always a chance of finding individuals in both the treatment and control groups. It rules out the possibility that some individuals with particular sets of observable characteristics are always found in one group and never in the other. For this reason, we propose that the first assessment to be conducted is to assess the overlap between the

treatment and control groups. If there are no problems in the degree of overlap, a straightforward approach is to assume that a regression model is a good approximation of the effect of the variables on the outcome and use multivariate regression (including a dummy variable to indicate whether individuals were treated or untreated). This assumes that the effect of the covariates on the outcome and the variance of the error term is the same for the treated and untreated groups. RA fits different regressions to the treated and control groups and hence offers additional flexibility. In the situation that a regression model is considered not to be a good approximation of the effect of the variables on the outcome (e.g. due to the parametric assumption imposed on the outcome variable), IPW, doubly robust methods or matching can be used. IPW makes no parametric assumptions on the distribution of the outcome. Instead, it estimates a weighted mean of the outcome in the treated and untreated groups, where each observation is weighted by the inverse of the probability of the individual belonging to each group. IPW can be combined with RA in what is termed 'doubly robust methods'.

Poor overlap can be improved with matching. Overlap should be assessed after matching. Trimming the sample to the region of overlap is a possibility for when the overlap is poor after trying different matching options. In doing so, the sample of individuals for which the treatment effect is calculated is redefined. Importantly, the estimates of the treatment effect refer to ATT rather than ATE if a large proportion of untreated individuals are removed from the analysis. Once the sample is trimmed to the region of overlap, the standard approaches for selection on observables can be used. Unadjusted or adjusted comparison of means (with multivariate regression) can be conducted if the overlap, post-matching, is satisfactory. If there are still small imbalances in the covariate distributions, a regression on the matched sample may be enough to control for those imbalances.

Figure 1: Proposed algorithm for selection of methods

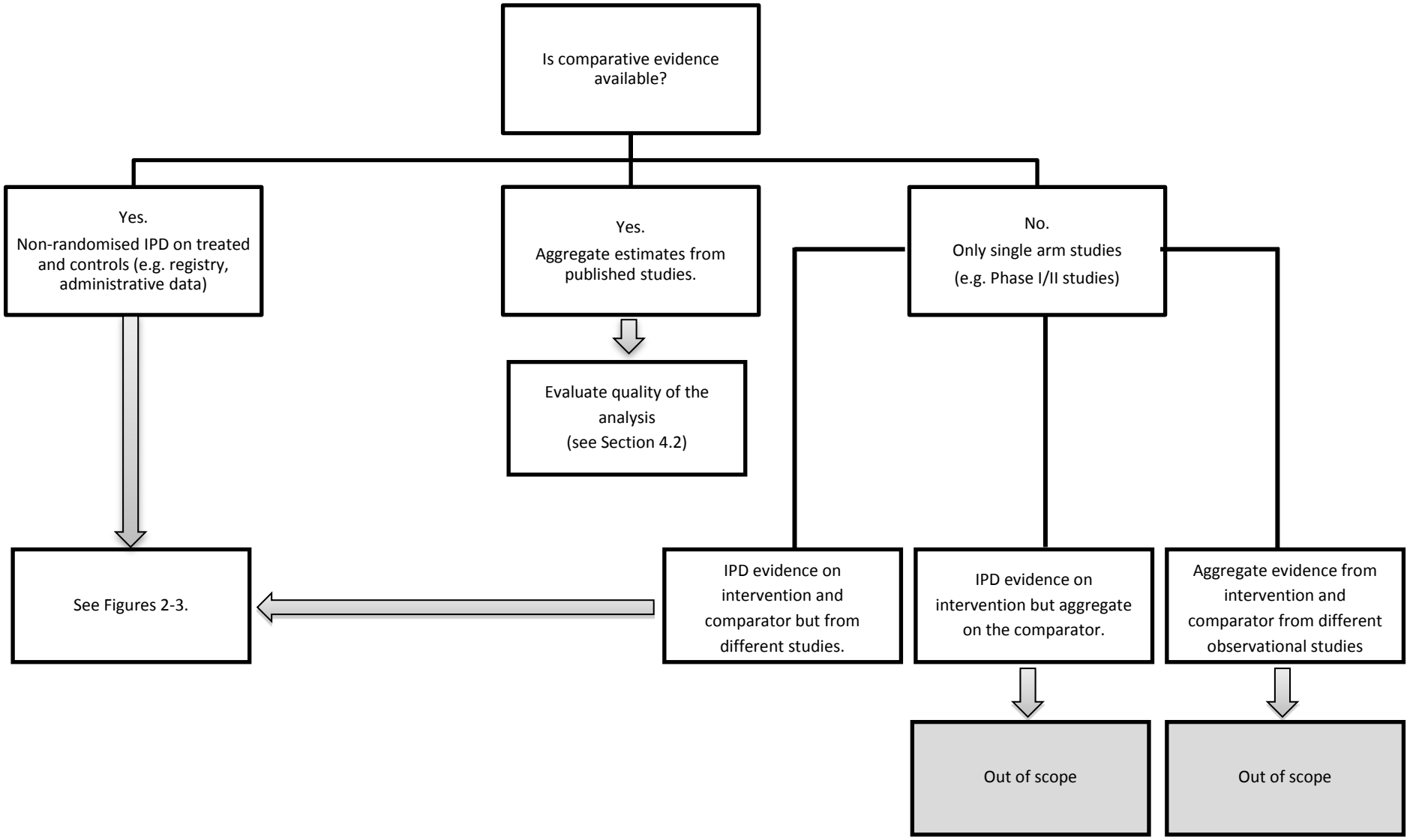


Figure 2: Continuation of proposed algorithm for selection of methods

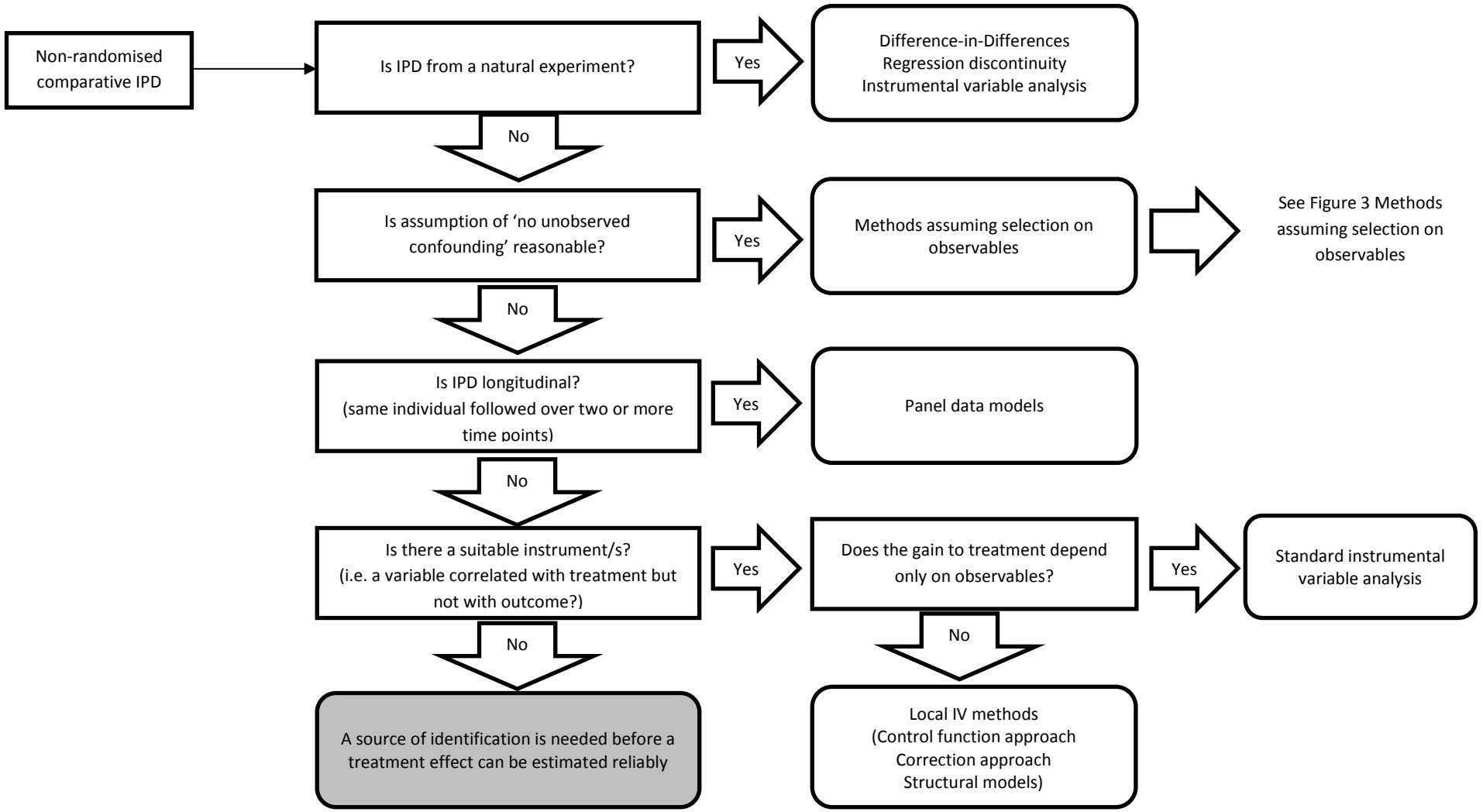
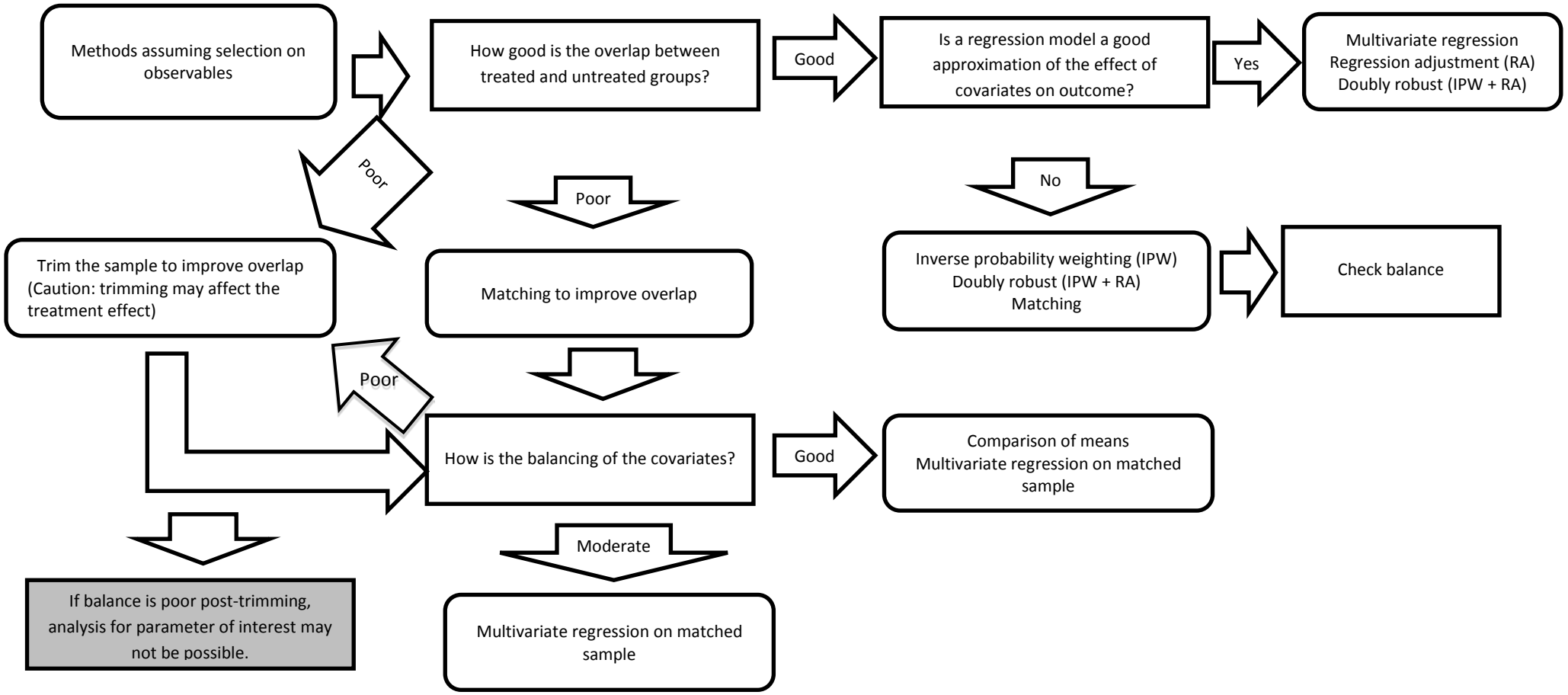


Figure 3: Continuation of proposed algorithm for selection of methods: Methods assuming selection on observables



4.2. HOW TO EVALUATE THE QUALITY OF AN ANALYSIS ON TREATMENT EFFECT USING NON-RANDOMISED DATA

In the context of TAs, companies and assessment groups may need to resort to estimates of treatment effect from published studies using non-randomised data. Therefore, it is important to understand how to critically appraise these studies to conclude whether the estimates are sufficiently robust to be used in the appraisal. Checklists can be a useful tool for critical appraisal. A checklist consists of a number of questions or tasks on a specific topic. Checklists can help ensure that all important issues are considered and improve consistency across different users, particularly for those with non-expert knowledge on the area.

4.2.1. Comparison of checklists: ISPOR, Kreif et al., GRACE and STROBE checklists

A pragmatic review of a selection of checklists on the use of non-randomised data was conducted to identify the key themes relating to the analysis of non-randomised data and to what extent the questions provide sufficient guidance for the critical appraisal of the methods. To our knowledge, there are three checklists on the use of non-randomised data in the context of cost-effectiveness analysis: the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Good Research Practices task force questionnaire to assess the relevance and credibility of prospective observational studies to inform healthcare decision making (ISPOR 2014 questionnaire),[55] the ISPOR checklist for retrospective database studies (ISPOR 2003 checklist)[56] and the checklist by Kreif *et al.* for critically appraising statistical methods to address selection bias in estimating incremental costs, effectiveness and cost-effectiveness (Kreif *et al.* checklist).[57] There are also checklists pertaining to the estimation of treatment effectiveness from non-randomised studies more generally, such as the Good ReseArch for Comparative Effectiveness (GRACE) checklist[58] and the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist.[59]

The questions in the checklists were extracted, compared and categorised to understand how the quality of the analyses was evaluated. Appendix 2 presents the comparison of checklists. Table 1 summarises the key themes emerging from the questions on the analytic methods and the corresponding questions from each checklist.

Table 1: Key themes on analytic methods from ISPOR, Kreif *et al.*, GRACE and STROBE checklists.

ISPOR 2003 checklist	ISPOR 2014 questionnaire	Kreif *et al.* checklist	GRACE checklist	STROBE checklist
Key theme 1: Minimising selection and confounding biases				
Control variables: if the goal of the study is to examine treatment effects, what methods have been used to control for other variables that may affect the outcome of interest?	1. Was there a thorough assessment of potential measured and unmeasured confounders?	1a. Did the study address the 'no unobserved confounding' assumption?	M3: Were important covariates, confounding and effect modifying variables taken into account in the design and/or analysis?	12. (a) Describe all statistical methods, including those used to control for confounding
Relevant variables: have the authors identified all variables hypothesised to influence the outcome of interest and included all available variables in their model?		1b. Did the study assess the assumption that the instrumental variable was valid?		
Key theme 2: Statistical equation for outcomes				
Statistical model: have the authors explained the rationale for the model/statistical method used?		3. Did the study assess the specification of the regression model for costs and health outcomes?		11. Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why
Testing statistical assumptions: do the authors investigate the validity of the statistical assumptions underlying their analysis?				
Model prediction: if the authors utilise multivariate statistical techniques in their analysis, do they discuss how well the model predicts what it is intended to predict?				
Key theme 3: Subgroups and interactions				
Multiple tests: if analyses of multiple groups are carried out, are the statistical tests adjusted to reflect this?	2. Were analyses of subgroups or interaction effects reported for comparison groups?			12. (b) Describe any methods used to examine subgroups and interactions
Key theme 4: Overlap (comparability) of treated and control groups				
		2. Did the study assess whether the distributions of the baseline covariates overlapped between the treatment groups?		12. (b) Case-control study—If applicable, explain how matching of cases and controls was addressed
		4. Was covariate balance assessed after applying a matching method?		
Key theme 3: Sensitivity analyses and uncertainty				
Statistics Influential cases: have the authors examined the sensitivity of the results to influential cases?	3. Were sensitivity analyses performed to assess the effect of key assumptions or definitions on outcomes?	5. Did the study consider structural uncertainty arising from the choice or specification of the statistical method for addressing selection bias?	M5: Were any meaningful analyses conducted to test key assumptions on which primary results are based?	12. (e) Describe any sensitivity analyses

Key theme 1 is on the analytic steps taken to minimise selection and confounding biases, both in terms of observed and unobserved confounders. All checklists included questions on this theme: e.g. "*Have the authors identified all variables hypothesized to influence the outcome of interest and included all available variables in their model?*" from the ISPOR 2003 checklist,[56] "*Were important covariates, confounding and effect modifying variables taken into account in the design and/or analysis?*" from the GRACE checklist.[58] The questions ask whether potential confounders have been identified and dealt with and provide some guidance on how to assess the methods were appropriate. For example, the Kreif *et al.* checklist suggests that the 'no unobserved confounding' assumption may be assessed with causal diagrams informed by external literature or in a discussion of external literature and expert opinion.[57] The ISPOR 2003 checklist indicates that two common approaches are stratification by different levels of confounding variables and multivariate statistic techniques. The GRACE checklist mentions that appropriate methods may include restriction, stratification, interaction terms, multivariate analysis, propensity score matching, instrumental variables or other approaches.

Key theme 2 is on whether the outcomes equation is appropriate. This includes questions on the rational for the model or statistical method used (e.g. "*Have the authors explained the rationale for the model/statistical method used?*" from the ISPOR 2003 checklist[56]) and the specification of the regression model (e.g. "*Did the study assess the specification of the regression model for costs and health outcomes?*" the Kreif *et al.* checklist[55]). These questions prompt the reader to question the assumptions underlying the method used. The ISPOR 2003 checklist exemplified the question with the issue of pooling individuals across centres without addressing the hierarchical nature of the data. Other examples include assuming a normal distribution for a non-normal dependent variable or assuming a linear relationship between the variables when the true relationship takes another format. These issues are relevant in any regression analysis and are discussed in detail in another DSU report.[21]

Key theme 3 is on subgroups and interactions, namely whether the study reports analyses of subgroups or interaction effects (such as in the ISPOR 2014 questionnaire[53] and STROBE checklist[57]) and whether statistical tests were adjusted for multiple testing (ISPOR 2003 checklist[54]). The ISPOR 2014 questionnaire explains that large observational studies have the potential to explore heterogeneous treatment effects (i.e. when the treatment effect is different

to subgroup populations) or effect modifiers (i.e. factors that interact and change the treatment effect). However, it cautions the reader to statistically significant subgroup effects when the effect in the entire population is not significant. This issue is also discussed in the ISPOR 2003 checklist, which recommends that multiple subgroup tests should be adjusted for.

Key theme 4 is on the comparability of treated and control groups. This is the focus of two questions of the Kreif *et al.* checklist: "*Did the study assess whether the distributions of the baseline covariates overlapped between the treatment groups?*" and "*Was covariate balance assessed after applying a matching method?*". Kreif *et al.* suggest that overlap can be assessed with histograms and standardised differences between groups. This is an important theme for the methods assuming no selection on unobservables. As discussed in Section 2, these methods assume overlap between treated and control groups. In other words, that for any combination of covariates, there is always the chance of observing individuals in the treated and control groups. If the overlap assumption does not hold, the estimated treatment effect may be biased or may only be applicable to the treated group (it is ATT rather than ATE).

Key theme 5 is on how uncertainty was addressed in the analysis. As with key theme 1 on selection and confounding biases, all checklists included questions about uncertainty. The ISPOR 2014 questionnaire, the GRACE and the STROBE checklists asked whether and how sensitivity analyses were conducted. The Kreif *et al.* checklist focussed on whether structural uncertainty arising from the choice and specification of the method had been considered. The ISPOR 2003 checklist included a question on whether the sensitivity of the results to outliers had been examined. Uncertainty is an important consideration in cost-effectiveness analysis (see Griffin *et al.*[60] for an extensive discussion). Parameter uncertainty refers to the distribution of potential realisations of the outcome and it is quantified in the standard error around the mean. Structural uncertainty relates to the assumptions required for the analysis, such as no selection on unobservables, correct model specification, etc. The ISPOR 2014 questionnaire recommends that the uncertainty analysis should consider different statistical approaches, the impact of removing outliers and the impact of potential unobserved confounders.[53] Kreif *et al.* suggest considering using methods based on different structural assumptions, doubly robust methods, different specifications of the outcomes equation, and exploring the effect of potential unmeasured confounders.[57]

The five checklists reviewed here included questions on the methods employed for the analysis of non-randomised data. The key themes emerging related to how selection and confounding biases were minimised, the correct specification of the outcomes equation, heterogeneity in treatment effect, the comparability of treated and control groups and the assessment of uncertainty. The key themes covered by all checklists were on selection and confounding biases and uncertainty. This suggests that they were unanimously considered to be important for the assessment of the quality of the analysis. There is some detail given on how to assess that the methods have been applied correctly. For example, the Kreif *et al.* checklist provides some guidance on how to assess the assumption of no selection on unobservables, whether the IV is valid, the correct specification of the regression model, covariate balance and how to explore structural uncertainty. However, there is little guidance on how to assess whether specific methods have been applied correctly. For example, whether the propensity score has been correctly estimated or whether the type of matching chosen is appropriate. An important limitation of this review is that it only included five checklists. There may be other checklists available that include detailed questions on how the methods were applied or that raise other important themes for consideration.

4.2.2. QuEENS : Quality of Effectiveness Estimates from Non-randomised Studies

The findings of the review of checklists motivated the development of a checklist to assess whether the methodology used to estimate treatment effect from non-randomised studies has been correctly applied. This Quality of Effectiveness Estimates from Non-randomised Studies (QuEENS) checklist was informed by the review of checklists in the earlier section, the methods review in Section 2 and the authors' experience in analysing non-randomised data. The objective of QuEENS is to help analysts unfamiliar with the methods for non-randomised data to critically appraise the quality of the analysis. QuEENS will also be useful for analysts in trying to select the right model or models to estimate the treatment effect of interest. It can be used on its own or as a complement to the checklists reviewed in Section 4.2.1, for example to help assess the application of a specific method.

Table 2 below presents the proposed checklist for the assessment of quality of effectiveness estimates from non-randomised studies (QuEENS). The supplementary notes below provide a useful complement to the table presenting the rationale for the questions in the checklist. QuEENS includes general questions on the analysis and specific questions relating to the

method employed in the study. Appendix 3 shows the application of the checklist to a published study.

Table 2: QuEENS: Checklist to assess the quality of effectiveness estimates from non-randomised studies

	Questions	Options		Comments
General issues	Q1: Have different methods been compared within the study?	(a) Yes		
		(b) Partially		
		(c) No		
	Q2: Have the results of the study been compared to others in the literature?	(a) Yes, compared to alternative methods using the same dataset		
		(b) Yes, compared to similar methods using other data sources		
		(c) Not compared – no other estimates found in the literature		
		(d) Not compared		
	Q3: Is there a discussion of what treatment effect is identified and of the assumptions needed?	(a) Discussion of effect and assumptions		
		(b) Discussion of effect but not the assumptions		
		(c) Discussion of the assumptions but not the effect		
		(d) No discussion of either		
	Q4: Is the model chosen consistent with the outcome variable if using a parametric method?	(a) Yes		
		(b) Unclear		
		(c) No		
	Q5: Were any checks conducted on the model specification?	(a) Yes, appropriate (detail which)		
		(b) Yes, but inappropriate or not enough		
		(c) No checks reported		
Methods assuming selection on observables	Q6: On selection: Is the assumption of selection on observables assessed?	(a) Yes, expert literature or opinion cited		
		(b) Yes, theoretical reasoning given.		
		(c) No		
	Q7: What checks were conducted to assess overlap?	(a) Yes, thorough checks		
		(b) Yes, minimum checks		
		(c) No checks reported.		
	Q8: Has balancing of the covariates been checked after matching and propensity score methods?	(a) Yes, thorough checks	Which ones?	
		(b) Yes, minimum checks		
		(c) No checks reported		

	Questions	Options		Comments
Methods using the propensity score	Q9: Is the propensity score function sufficiently flexible?	(a) Yes, includes interactions or different functions of the covariates		
		(b) Yes, flexible due to the semi-parametric/non-parametric specification		
		(c) Unlikely to be flexible enough		
		(d) Unclear or not reported.		
	Q10: Are potential IVs excluded from the set of conditioning variables?	(a) Yes		
		(b) Some variables might present a problem		
		(c) IV clearly included		
Matching methods	Q11: Data quality: Are there data quality issues?	(a) Data and definitions comparable for treated and control groups	Yes	
			No	
			Unclear or not reported	
		(b) Treated and controls come from the same area or environment	Yes	
			No	
			Unclear or not reported	
		(c) Rich set of variables used for matching	Yes, available and used	
			Not available or not used	
		(d) Reasonable sample sizes	Yes, likely	
			No	
	Q12: For Nearest Neighbour: Has bias adjustment been conducted if more than one variable was included when matching on covariates?	(a) Yes		
		(b) No		
	Q13: Is the choice of replacement (with/without) reasonable?	(a) Yes		
		(b) Likely		
		(c) No		
	Q14: Is the choice of the number of matches/calliper matching/radius matching reasonable?	(a) Yes		
		(b) Likely		
		(c) No		
IV methods	Q15: Is the instrument well justified? (i.e. eligibility in programme participation (reason), natural experiment, theoretically sensible, fitted propensity scores)	(a) Yes, theoretically		
		(b) Yes, citing expert literature		
		(c) No		
	Q16: Is the sample size relatively large?	(a) Yes		
		(b) No		
	Q17: If more than one IV, is the test of over-identifying restrictions reported?	(a) Yes		
		(b) No		

	Questions	Options		Comments
	Q18: Is a weak instrument(s) test reported?	(a) Yes		
		(b) No		
Difference in Differences (standard)	Q19: Does the intervention generate exogenous variation? (not applicable if natural experiment)	(a) Yes, highly likely		
		(b) Unlikely		
		(c) Not applicable		
	Q21: Is the assumption of common trends across groups reasonable?	(a) Yes, highly likely		
		(b) Unlikely		
	Q22: Is it reasonable to assume that there is no selection of unobserved temporary individual specific shocks?	(a) Yes, highly likely		
		(b) Unlikely		
	Q23: Is the assumption of no systematic composition changes within each group reasonable? (applicable with repeated cross-sections, not with longitudinal data)	(a) Yes, highly likely		
		(b) No, unlikely		
		(c) Not applicable		
Regression discontinuity design	Q24: Is the sample size relatively large?	(a) Yes		
		(b) No		
	Q25: Is the assumption that individuals are not able to affect the instrument to change the likelihood of participation reasonable?	(a) Yes, highly likely		
		(b) Unlikely		

A fundamental requirement for choosing an appropriate method or methods to estimate the treatment effect in non-randomised studies is to have a clear understanding of the process of treatment assignment. Obviously this is case specific and therefore it is expected that any study using non-randomised data will require a detailed discussion about the nature of the data, how it has been collected/generated and the mechanism of treatment assignment. This section aims to assist the reader make an informed assessment about the assumptions that are likely to hold in each particular case. It gives comprehensive details for each question in the checklist and provides details of tests, graphical aids, rationale and other elements of reporting that one should expect to see in an applied research piece to justify the appropriateness of the models and therefore to substantiate the results.

General issues (Questions 1-5)

This section considers issues that are of general applicability to any study in this area.

Question 1. Have different methods been compared within the study?

It is reasonable for any study attempting to estimate treatment effects to implement a number of methods based on different assumptions. This could be used to gauge the sensitivity of the results to the assumptions underpinning the models. At the same time, adopting a number of approaches forces the analyst to think about the assumptions embedded in each of the methods and their plausibility and helps focus on those of most importance. However, different methods might be estimating different treatment parameters and therefore different numerical parameters might be the result of this. The possible answers to the above question are as follows:

(a) Yes

Results from methods which assume selection on observables are contrasted with other methods, including those assuming selection on unobservables.

(b) Partially

Results from different methods are contrasted but all the methods rely on the same assumption about selection, either selection on observables or selection on unobservables.

(c) No.

Question 2. Have the results of the study been compared to others in the literature?

Similar to Question 1 above, a study should compare its results to those found in the literature. Given that they would relate to different methods and/or different datasets, one would expect differences in the results but consistency between them (or inconsistencies that are easily explained) will give credibility to the results. The possible answers here are:

(a) Yes, compared to alternative methods using the same dataset.

(b) Yes, compared to similar methods using other data sources.

(c) Not compared – no other estimates found in the literature. This option should be selected when there is an indication that a search was conducted in the literature but no other related estimates were found.

(d) Not compared.

Question 3. Is there a discussion of what treatment effect is identified and of the assumptions needed?

Usually the parameter of interest in economic evaluations for NICE is the ATE but in some cases it might be the ATT. The parameter of interest in the analysis should match the

parameter of interest in the economic evaluation. In Section 2, the different types of treatment effects that can be identified were discussed and were related to the different approaches and their assumptions. Any study should show an awareness of this issue. For example, if one is willing to make the assumption of homogeneity in the treatment effect, then it is straightforward to identify the ATE. With heterogeneity, the ATT might be identified under weak assumptions. However, the ATE may need a much more stringent set of assumptions. If the parameter of interest is the ATT, this is not a problem. A good study should justify how the estimated treatment effect related to the treatment effect of interest, together with their underpinning assumptions. The possible options to be selected are as follows:

(a) Discussion of effect and assumptions.
(b) Discussion of effect but not the assumptions.
(c) Discussion of the assumptions but not the effect.
(d) No discussion of either.

Question 4: Is the model chosen consistent with the outcome variable if using a parametric method?

The distribution of the outcome variable should inform the choice of the type of regression model to use. For example probit/logit models can be used with binary outcomes, generalised linear models can be very useful in cases where the data is highly skewed, etc. The possible options to be selected are as follows:

(a) Yes
(b) Unclear
(c) No

Question 5: Were any checks conducted on the model specification?

Specification checks should be conducted on the models. The appropriate checks will depend on the model used. For example linear regression models can be assessed using plots of the residuals or more formally using misspecification, heteroskedasticity, autocorrelation, normality, etc. tests based on the residuals; if using kernel regression or matching it is important to check the sensitivity of the results to the choice of bandwidth and matching algorithm respectively (see Wooldridge, Jones and Rice, Kreif *et al.*[14,15,57]). The possible options to be selected are as follows:

(a) Yes, appropriate (detail which)
(b) Yes, but inappropriate or not enough

(c) No checks reported

Methods assuming selection on observables (Questions 6-8)

Question 6. On selection: Is the assumption of selection on observables assessed?

The methods presented in Section 2.2 are based on the assumption that selection is on observables. Strictly speaking, selection is on both observables and unobservables but the unobservables are not correlated with the outcomes and thus, their presence does not induce confounding. This assumption is often controversial and cannot be tested directly although placebo tests can sometimes be used. A convincing argument should put forward to substantiate the claim that the selected variables are sufficient and, once used in the analysis, there are no remaining unobserved variables affecting both the treatment and the outcome. The following options are available:

(a) Yes, expert literature/opinion cited. The analyst justifies the assumption with reference to *a priori* knowledge in the expert literature or if this is lacking with reference to expert opinion. Sometimes it is possible to assess this assumption indirectly by testing if a treatment effect is zero when it is known that it is. For example, if there is access to two different control groups, one can check that the treatment effect is zero between the two groups or one can use a variable known not to have an effect to estimate the treatment effect.

(b) Yes, theoretical reasoning given. The analyst justifies the assumption with a sensible theoretical argument but does not refer to the literature.

(c) No.

Question 7. What checks were conducted to assess overlap?

All methods assuming selection on unobservables rely on good overlap in the distribution of the covariates between the treatment and control groups. Even if ignorability holds, the results will be suspect if there is lack of overlap between the treatment and control groups. Lack of overlap implies that regression estimates extrapolate to regions well outside the sample, might cause instability in estimates using IPW and call into question matching estimates of the average treatment effect as it will not be possible to find matches for some individuals.

(a) Yes, thorough checks. As a starting point, it is useful to report normalised differences in covariates for the treatment and the control groups to check if overlap is a problem. Normalised differences above 0.25 have been suggested as signalling problems with

overlap. It is important to emphasise that normalised differences are different from the usual t-statistics of the difference in means between the treatment and control groups. Looking at one covariate at a time and focusing only on one moment (the mean) in its distribution is insufficient. Other more thorough checks include comparing histograms or kernel plots of the covariates for the treatment and the control groups, quantile-quantile (QQ) plots, higher moments and cross moments of covariate distributions. If there are many covariates or the propensity score is estimated as part of the model, a better alternative is to present distributions of the propensity score by treatment group because we are trying to assess if there are any areas where the density of the covariates is zero for one group and non-zero for the other. Note that the overlap in the covariates will most likely be assessed as part of a nonparametric regression method for example.

(b) Yes, minimum checks. These include normalised differences at the very least and perhaps some but not all of the additional checks reported in (a).

(c) No checks reported.

Question 8: Has balancing of the covariates been checked after matching and propensity score methods?

Matching and propensity score methods should achieve balancing of the covariates.

(a) Yes, minimum checks. The analyst can use normalised differences appropriate for each methods in covariates for the treatment and the control groups or weighted normalised differences in the case of IPW as in Austin.[45]

(b) Yes, more thorough checks. Other more thorough checks include comparing histograms or kernel plots of the covariates for the treatment and the control groups, or if matching on the propensity score comparing distributions of the propensity score by treatment.

(c) No checks reported.

Methods using the propensity score (Questions 9-10)

Question 9: Is the propensity score function sufficiently flexible?

It has been suggested that the propensity score function needs to be sufficiently flexible and therefore should include not just the variables in levels but also squares and interactions. Clearly, the flexibility will depend on the size of the dataset. One can also use

semiparametric/non-parametric functions to model the propensity score. The available choices for this question are:

(a) Yes, includes interactions or different functions of the covariates
(b) Yes, flexible due to semiparametric/non-parametric specification
(c) Unlikely to be flexible enough
(d) Unclear or not reported

Question 10: Are potential IVs excluded from the set of conditioning variables?

Variables that should be included in the conditioning set are variables measured before the assignment to treatment takes place, including past outcomes. Variables that are potential IVs should not be included because they have been shown to increase the bias in matching type estimators unless they are exogenous. Even in this case, when they do not cause a bias, they will increase the asymptotic variance of the estimate. The available choices are:

(a) Yes
(b) Some variables might present a problem
(c) IV clearly included

Matching methods (Questions 11-14)

Question 11: Are there data quality issues?

An important issue in matching is the quality of the data. For the treatment effect calculated using matching to be convincing, the data and the definitions for the treated and control groups must be comparable. The assumption of no unobserved confounders remaining which affect both the treatment and the outcome is more compelling if the treated and controls come from the same, or at least very similar, environment. It is also important that the dataset includes a good number of variables that can be used for matching and that the sample sizes before matching are big enough so there are plenty of potential matches. Accordingly, the following subcategories are available:

(a) Data and definitions comparable for treated and control groups: Yes/No/Unclear or not reported.
(b) Treated and control come from the same area or environment: Yes/No/Unclear or not reported.
(c) Rich set of variables: Yes, available and used/Not available or not used.
(d) Reasonable sample sizes: Yes, likely/ No.

Question 12: For Nearest Neighbour matching: Has bias adjustment been conducted if more than one variable was included?

Abadie and Imbens[61,62] showed that the estimator obtained using Nearest Neighbour matching is biased if matching on more than one continuous covariate and proposed a bias adjustment. Imbens and Wooldrige[13] highlight the cases under which the bias will be small in practice.

(a) Yes

(b) No

Question 13: Is the choice of replacement (with/without) reasonable?

Matching without replacement if the control group is small might result in bad matches which increase the bias of the estimator. Matching with replacement might result in the same individuals in the control group being matched to in areas of the propensity score where there are many more treated observations than controls. This means that some untreated individuals may be matched repeatedly. One of the following options should be selected:

(a) Yes

(b) Likely

(c) No

Question 14: Is the choice of the number of matches/caliper matching/radius matching reasonable?

There is a trade-off between bias and variance which the analyst needs to take into account. Note that this is a subjective decision and there is not much known about, for example, how to select the number of matches.

(a) Yes

(b) Likely

(c) No

IV methods (Questions 15-18)

Question 15: Is the instrument well justified?

An IV variable needs to affect the treatment directly but the outcome only indirectly through its effect on the treatment. This exclusion restriction is key but cannot be tested directly. If there is more than one IV, one can test over-identifying restrictions (see Question 17) but in most cases one needs to rely on the published literature and expert opinion.

(a) Yes, theoretically

(b) Yes, citing expert literature

(c) No

Question 16: Is the sample size relatively large?

IV methods are biased on finite samples but they are consistent in large samples. Therefore it is important that they are used in relatively large datasets.

(a) Yes

(b) No

Question 17: If more than one IV, is the test of over-identifying restrictions reported?

A test of over-identifying restrictions is essentially a test of instrument validity and should be reported whenever there are more instruments than endogenous variables. If the number of endogenous variables is the same as the number of instruments, one can always create additional instruments by interacting the IV with other covariates in the model. Note that rejection of the hypothesis could be due to a failure of the instrument but also to model misspecification.

(a) Yes

(b) No

Question 18: Is a weak instrument(s) test reported?

Weak instruments lead to an increase in the bias of the estimator. Simple correlations or partial correlations can be used in the first instance. More formal tests such as that reported in Cragg and Donald[63] could also be used.

(a) Yes

(b) No

Difference in Differences (Questions 20-23)

The following sets of questions relate to assumptions that are untestable and it is therefore important that the analyst justifies them with reference to the published literature or expert opinion.

Question 20: Does the intervention generate exogenous variation? (not applicable if natural experiment)

The DiD approach makes use of interventions or events which induce random assignment of the individual to the treatment and control groups or at least random eligibility. This is similar to the exogenous variation in the treatment variable achieved by randomisation. In general, this is not applicable for natural experiments although it is always appropriate to assess if the natural experiment generated exogenous variation.

(a) Yes, highly likely
(b) Unlikely
(c) Not applicable

Question 21: Is the assumption of common trends across groups reasonable?

Differential trends might arise if for example, the treatment and control groups are based in different areas with different trends in the outcomes, or when external shocks to the outcome happen at different time points. The differential trend adjusted DID estimator can be used is the trends might not be the same and the analyst has access to historical data (see Section 2.5.1).

(a) Yes, highly likely
(b) Unlikely

Question 22: Is it reasonable to assume that there is no selection of unobserved temporary individual specific shocks?

This question relates to the Ashenfelter's dip discussed in the previous section. If individuals are able to change their behaviour before the timing of the treatment to manipulate their probability of getting the treatment, the DID method will not be able to identify the correct treatment effect.

(a) Yes, highly likely
(b) Unlikely

Question 23: Is the assumption of no systematic composition changes within each group reasonable? (applicable with repeated cross-sections, not with longitudinal data)

The DiD method is able to remove the unobserved individual effect using repeated cross-sections only if there are no composition changes in the groups so that the average unobserved individual effect remains the same before and after the treatment or intervention.

(a) Yes, highly likely
(b) Unlikely
(c) Not applicable

Regression Discontinuity Design (Questions 24-25)

Question 24: Is the sample size relatively large?

In common with IV methods, the regression discontinuity design identifies a local parameter and therefore the estimates may not be very precise if the sample size is small.

(a) Yes
(b) No

Question 25: Is the assumption that individuals are not able to affect the instrument to change the likelihood of participation reasonable?

The regression discontinuity design will in general not be able to identify the required treatment effect if individuals are able to manipulate the instrument to increase/decrease their likelihood of participation. In this case, individuals below and above the threshold are different in terms of the unobservables.

(a) Yes
(b) No

5. DISCUSSION

5.1. SUMMARY OF FINDINGS

The estimation of treatment effect from non-randomised data poses challenges over and above RCTs. Non-random treatment assignment can lead to differences in the factors affecting individuals in the treated and control groups. Confounding bias arises if there are differences in factors that affect outcomes. There are a number of methods that have been developed to minimise the risk of confounding bias. These methods can be broadly classified as those controlling for observed confounders (i.e. methods assuming no selection on unobservables) and those that can control for unobserved confounders (methods assuming selection on unobservables). Section 2 presents a review of the most common and straightforward to apply methods in both categories. The methods that assume selection on observables require a good degree of overlap between treated and control groups; this means that there is a chance of observing the treated and control groups for any combination of covariates. These methods include: multivariate regression, regression adjustment, inverse probability weighting, doubly robust, regression on the propensity score, matching, and regression on a matched sample. The methods assuming selection on unobservables can obtain an unbiased estimate of treatment effect when the treated and control groups are different in unobserved factors that affect the outcome. Section 2 reviewed some of the techniques to deal with this issue, in particular instrumental variable methods and briefly covered panel data models. Other methods reviewed that can handle selection on unobservables are those making use of natural experiments, namely differences-in-differences and regression discontinuity.

The review of NICE TAs using non-randomised data to obtain estimates of treatment effect illustrated the types of data used, which included comparative IPD of the different interventions and comparators, IPD only for the intervention, and aggregate data (either as an estimate of treatment effect obtained from a published study or aggregate estimates of outcome for the interventions or comparators). The TAs that used comparative IPD used multivariate regression and propensity score matching to adjust for (observed) differences between the intervention groups. Most of the TAs identified by the review used aggregate data from published studies and obtained estimates of treatment effect unadjusted for potential confounders. The review suggested that these analyses could have been improved in

many ways, which stress the need for guidance in the analysis, interpretation and critical appraisal on the use of non-randomised evidence for TA.

Section 4 provides guidance on the use of non-randomised comparative IPD to estimate treatment effect and on how to critically appraise studies that have reported estimates of treatment effect from non-randomised data. The algorithm for method selection suggests a series of steps on how to choose the appropriate method given the available data. The review of checklists indicates the key themes of concern in the analysis of non-randomised data, namely how selection and confounding biases were minimised, the correct specification of the outcomes equation, heterogeneity in treatment effect, the comparability of treated and control groups and the assessment of uncertainty. However, the available checklists do not include questions specific to the implementation of each of the methods. Therefore, analysts unfamiliar with the area may find it difficult to critically appraise a study using non-randomised data. For this reason, a new checklist, QuEENS, was developed to include questions and supplementary notes on the implementation of each method. QuEENS should be useful to both critically appraise published studies and to help guide the implementation of the different methods. The application of QuEENS is exemplified with a non-randomised study used in a previous TA (Edidin *et al.*,[64] used in TA279).

The issues around the use of non-randomised data to estimate treatment effect may increase the complexity of NICE TAs. This TSD can help companies, assessment groups and appraisal committees to become aware of the challenges in using non-randomised data and the possible solutions.

5.2. FINAL RECOMMENDATIONS FOR ANALYSIS

There are a number of important considerations for the analysis of non-randomised comparative IPD that arise from the reviews and development of new tools for implementation and critical appraisal of methods (algorithm and QuEENS checklist):

- Consider whether the data available is appropriate to answer the decision problem and to inform the estimation of the parameter of interest for the cost-effectiveness model, namely in terms of patient population, interventions, comparators, setting and available data on potential confounders. Consider the parameter of interest (ATE, ATT, other) and how it can be estimated from the available data.

- Justify the choice of method for the base-case. Demonstrate that the variables involved in treatment assignment are included in the dataset, if modelling treatment assignment based on observables. Alternatively, if the method can handle selection on unobservables, justify how the implementation of the method can ensure an unbiased estimate of treatment effect.
- Justify the implementation of the method, namely the variables included in the outcome equation, the functional form of the model, parametric assumptions. Test the assumptions with model specification tests and sensitivity analysis using alternative specifications. Discuss how the results of the tests support the implementation of the method and test alternative specifications in the sensitivity analysis.
- As with an RCT, the analyst must state upfront the assumed mechanism of causality, i.e. how and why is the intervention expected to affect the outcome. Interpret the results in light of current knowledge, being mindful of counter-intuitive associations between covariates and the outcome of interest (e.g. comorbidities known to increase risk of death appearing to have a protective effect). Counter-intuitive results may suggest the presence of omitted variable bias and hence unobserved confounding.
- Conduct pre-planned sensitivity analysis with different methods (e.g. matching vs multivariate regression) and different implementations of the same method (e.g. with a different set of variables included in the matching process, inclusion of interactions or polynomial terms). Discuss the relative plausibility of the alternative approaches.
- Ensure transparent and comprehensive reporting of the all analyses conducted. Explain the assumptions of the methods and implications of the results in such a way that clinical experts can understand the analysis and validate the plausibility of the results.

5.3. STRENGTHS AND LIMITATIONS

This TSD provides an overview of the most commonly used methods to handle non-randomised data, proposes an algorithm for method selection, a checklist for quality assessment and exemplifies its application to a published study that has been previously used in a NICE TA. The review of NICE TAs illustrates how non-randomised data has been used to estimate treatment effect. The review of methods examines each in turn, discusses its assumptions in a narrative format aimed at applied analysts and indicates the most relevant references where the interested reader can access full technical details. The algorithm for

method selection proposes a structured approach to the analysis of non-randomised data whilst emphasising the assumptions underpinning each method. The checklist for critical appraisal suggests a sequence of questions, some general and some specific to the methods employed, to stimulate critical thinking and facilitate the appraisal. These tools, in combination with the good practice recommendations above, should help improve the quality of the analysis, reporting, critical appraisal and interpretation for decision making.

This TSD aimed to provide practical guidance on the methods that are straightforward to apply and are most commonly used in statistical analysis and econometrics of non-randomised data. The selection of methods was based on the main methods proposed in review articles and textbooks for non-randomised data, together with the authors' experience in the area. There are more sophisticated methods which are beyond the scope of this report. In addition, the literature in this area is advancing rapidly and generalisations and improvements of well-established methods as well as new methods are being continuously developed.

This TSD was restricted to methods to handle non-randomised *comparative* IPD and excluded IPD relating to a single group (i.e. non comparative) or situations where only aggregate estimates from different studies are available for each treatment group. The restriction in scope was due to the differences between the methods. The methods for uncontrolled IPD need to consider not only how to ensure comparability between treatment groups but also how to control for time trend and natural history of the disease. Alternatively, IPD or aggregate data from a control group external to the study needs to be identified. Situations in which only aggregate data from different studies are available create additional complications. Not only it is very difficult to adjust aggregate estimates from different treatment groups from different studies to account for different distributions of prognostic covariates but also it may be virtually impossible to assess whether the populations are comparable, the mechanisms of selection bias, outcomes assessment and drop-out. For these reasons, these two types of non-RCT evidence may be more adequately examined in a future TSD.

The focus was on the analytic methods to estimate treatment effect using non-randomised comparative data. Issues around data quality, study design or reporting were not considered. Unlike RCTs, non-randomised studies often make use of secondary databases that were

created for administrative (e.g. billing and record keeping) purposes rather than for research. Therefore, these data may be at a greater risk of errors compared to data collected in a controlled study, or may omit important covariates that would normally be collected in prospective research-based non-randomised studies. In addition, secondary databases may be limited to the individuals who seek care or who have access to the specific healthcare system. While in the UK the full population has access to the UK National Health Services, other countries may be different (e.g. US), which may have implications for the validity of the data. Study design is another important aspect of using non-randomised data. As discussed in Section 1, there is a wide range of study designs depending on the context and data availability. Different study designs may have implications for the analysis stage, namely restricting the set of methods that are applicable. Reporting was briefly mentioned in the recommendations (See Section 5.2). However, it may warrant being the focus of a future TSD since good quality and transparent reporting is essential for the critical appraisal of a study.

This TSD did not discuss the application of these estimates in decision model. Estimates of treatment effect from non-randomised data are subject to additional uncertainty compared to estimates from RCTs. The parameter uncertainty captured in the standard error around the mean estimates can be incorporated in the decision model with probabilistic sensitivity analysis. However, the structural uncertainty is more difficult to quantify. There is some structural uncertainty associated with the statistical method and its implementation. Importantly, there is structural uncertainty inherent to making inferences from non-randomised data. There is little guidance on how to incorporate structural uncertainty in decision analytic modelling.

The algorithm for method selection and the QuEENS checklist are a first step towards guidance on the analysis of non-randomised data. Both tools were informed by the authors' experience in non-randomised data and in NICE TA. However, they require further work, namely validation and additional testing with a variety of users to ensure they are useful and provide consistent results.

5.4. AREAS FOR FUTURE RESEARCH

There are a number of areas for future work that emerge and may warrant additional research or reviewing in future TSDs:

- Methods for non-comparative IPD. The review of NICE TAs in Section 3 indicated that there are situations where non-comparative IPD is the sole source of treatment effect. This is likely to become more frequent because the European Medicines Agency (EMA) has started to accept non-comparative IPD for regulatory approval under the conditional marketing authorisation process. Consequently, there may be a need for guidance on how to use non-comparative IPD for NICE TA.
- Methods for aggregate data. The review of NICE TAs also showed that companies and assessment groups may not have access to IPD or to estimates of treatment effect from good quality non-randomised studies. Companies and assessment groups may only have aggregate estimates of outcomes for the interventions and comparators. Using these aggregate outcomes without adjustment may provide a biased estimate of treatment effect because there may be differences in confounders between treatment groups. However, it is difficult to adjust aggregate estimates without access to IPD. A related issue is disconnected networks. Disconnected networks occur when there is no RCT that links two strands of the evidence synthesis network. Therefore, randomisation may need to be broken to enable the comparison of interest. The challenges related with the use of aggregate data may warrant additional guidance and research.
- Methods to combine IPD from randomised and non-randomised studies. There may be situations where IPD is available from both randomised and non-randomised studies for the same comparison. The different designs may be advantageous to improve the internal and external validity of the estimates. However, there may be challenges in the analysis and interpretation of the results.
- Non-randomised study designs. As non-randomised data becomes more frequent in NICE TAs, there may be scope for guidance on the appropriate study design to minimise bias and ensure reliable estimates of treatment effect. A related issue are novel trial designs, such as adaptive trials, that mix characteristics of randomised and non-randomised data.
- Structural uncertainty associated with non-randomised studies. The parameter uncertainty around the mean estimate of effect does not capture the uncertainty associated with the risk of bias of non-randomised data. Future research should consider how to characterise and incorporate this structural uncertainty in the decision modelling and value of information framework.

6. REFERENCES

1. Deeks J.J., Dinnes J., D'Amico R., Sowden A.J., Sakarovitch C., Song F. et al. Evaluating non-randomised intervention studies. *Health Technology Assessment* 2003; 7(27) http://www.journalslibrary.nihr.ac.uk/__data/assets/pdf_file/0007/64933/FullReport-hta7270.pdf

2. National Institute for health and Clinical Excellence. Guide to the methods of technology appraisal. 2013; http://publications.nice.org.uk/guide-to-the-methods-of-technology-appraisal-2013-pmg9/foreword

3. Rubin D.B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; 66(5):688.

4. Neyman J. Statistical problems in agricultural experiments. *Journal of the Royal Statistical Society* 1923; II((Supplement)):107-180.

5. Roy A.D. Some thoughts on the distribution of earnings. *Oxford Economic Papers* 1951; 3(2):135-146.

6. Heckman J.J. Econometric causality. *International Statistical Review* 2008; 76(1):1-27.

7. McClellan M., McNeill B.J., Newhouse J.P. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *JAMA* 1994; 272(11):859-866.

8. Imbens G.W., Angrist J.D. Identification and estimation of local average treatment effects. *Econometrica: Journal of the Econometric Society* 1994;467-475.

9. Heckman J.J., Vytlacil E.J. Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. *Handbook of Econometrics* 2007; 6:4875-5143.

10. Basu A., Heckman J.J., Navarro–Lozano S., Urzua S. Use of instrumental variables in the presence of heterogeneity and self–selection: an application to treatments of breast cancer patients. *Health Economics* 2007; 16(11):1133-1157.

11. Basu, A. Economics of individualization in comparative effectiveness research and a basis for a patient-centered health care. *Journal of Health Economics* 2011; 30(3):549-559.

12. Blundell R., Dias M.C. Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources* 2009; 44(3):565-640.

13. Imbens G.M., Wooldridge J.M. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 2009; 47(1):5-86.

14. Wooldridge J.M. Econometric analysis of cross section and panel data. 2010.

15. Jones Andrew M., Rice N. Econometric evaluation of health policies. 2011.

16. Nichols A. Causal inference with observational data. *Stata Journal* 2007; 7(4):507.

17. Nichols A. Erratum and discussion of propensity-score reweighting. *Stata Journal* 2008; 8(4):532-539.

18. Jones A.M. Identification of treatment effects in Health Economics. *Health Economics* 2007; 16(11):1127-1131.

19. Galiani S., Gertler P., Schargrodsky E. Water for life: The impact of the privatization of water services on child morality. *Journal of Political Economy* 2005; 113:83-120.

20. Jackson C., Thompson S.G., Sharples L.D. Accounting for uncertainty in health economic decision models by using model averaging. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2009; 172:383-404.

21. Kearns B., Ara R., Wailoo A., Manca A., Alava M.H., Abrams K. et al. Good practice guidelines for the use of statistical regression models in economic evaluations. *PharmacoEconomics* 2013; 31(8):643-652.

22. Nixon R.M., Thompson S.G. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Economics* 2005; 14(12):1217-1229.

23. Basu A., Manca A. Regression estimators for generic health-related Quality of Life and Quality-Adjusted Life Years. *Medical Decision Making* 2012; 32(1):56-69.

24. Hernández Alava M., Wailoo A., Wolfe F., Michaud K. A comparison of direct and indirect methods for the estimation of health utilities from clinical outcomes. *Medical Decision Making* 2014; 34(7):919-930.

25. Jones A.M., Lomas J., Rice N. Applying beta-type size distributions to healthcare cost regressions. *Journal of Applied Econometrics* 2014; 29(4):649-670.

26. Wooldridge J. Should instrumental variables be used as matching variables. 2009.

27. Rosenbaum P.R., Rubin D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; 70(1):41-55.

28. Crump R.K., Hotz V.J., Imbens G.W., Mitnik O.A. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 2009;asn055.

29. StataCorp. Stata 13 Base Reference Manual. 2013.

30. Radice R., Ramsahai R., Grieve R., Kreif N., Sadique Z., Sekhon J.S. Evaluating treatment effectiveness in patient subgroups: A comparison of propensity score methods with an automated matching approach. *International Journal of Biostatistics* 2012; 8(1):1557-4679.

31. Kreif N., Grieve R., Radice R., Sadique Z., Ramsahai R., Sekhon J.S. Methods for estimating subgroup effects in cost-effectiveness analyses that use observational data. *Medical Decision Making* 2012; 32(6):750-763.

32. Kreif N., Grieve R., Radice R., Sekhon J.S. Regression-adjusted matching and double-robust methods for estimating average treatment effects in health economic evaluation. *Health Services and Outcomes Research Methodology* 2013; 13(2-4):174-202.

33. Kang J.D.Y., Schafer J.L. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 2007; 22:523-539.

34. Robins J., Sued M., Lei-Gomez Q., Rotnitzky A. Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science* 2007; 22:544-559.

35. Kreif N., Gruber S., Radice R., Grieve R., Sekhon J.S. Evaluating treatment effectiveness under model misspecification: A comparison of targeted maximum likelihood estimation with bias-corrected matching. *Statistical Methods in Medical Research* 2014; doi: 10.1177/0962280214521341.http://smm.sagepub.com/content/early/2014/04/21/0962280214521341

36. Petersen M.L., Porter K., Gruber S., Wang Y., van der Lann M.J. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research* 2010; doi: 10.1177/0962280210386207.http://smm.sagepub.com/content/early/2010/10/27/0962280210386207.abstract

37. Mitra N., Indurkhya A. A propensity score approach to estimating the cost-effectiveness of medical therapies from observational data. *Health Economics* 2005; 14(8):805-815.

38. Manca A., Austin P.C. Using propensity score methods to analyse individual patient level cost effectiveness data from observational studies. *The University of York: Health Economics and Data Group Working Paper* 2008; 8(20) http://www.york.ac.uk/res/herc/documents/wp/08_20.pdf.

39. Abadie A., Imbens G.W. Matching on the estimated propensity score. *Harvard University and National Bureau of Economic Research* 2012. Accessed 7-7-2014 http://www.hks.harvard.edu/fs/aabadie/pscore.pdf

40. StataCorp. Stata Statistical Software: Release 13. 2013.

41. Lunt M. Selecting an appropriate caliper can be essential for achieving good balance with propensity score matching. *American Journal of Epidemiology* 2014; 179(2):226-235.

42. Diamond A., Sekhon J.S. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics* 2013; 95(3):932-945.

43. Stuart E.A. Matching methods for causal inference: A review and a look forward. *Stat Sci* 2010; 25(1):1-21.PM:20871802

44. Caliendo M., Kopening S. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys* 2008; 22(1):31-72.

45. Austin P.C. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Med Decis Making* 2009; 29(6):661-677.PM:19684288

46. Sekhon J.S., Grieve R. A matching method for improving covariate balance in cost-effectiveness analyses. *Health Economics* 2012; 21(6):695-714.

47. Basu A., Chan K.C.G. Can we make smart choices between OLS versus contaminated IV estimators? *Health Economics* 2014; 23(4):462-472.

48. Hausman J.A. Specification tests in econometrics. *Econometrica: Journal of the Econometric Society* 1978;1251-1271.

49. Craig P., Cooper C., Gunnell D., Haw S., Lawson K., Macintyre S. et al. Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. *J Epidemiol Community Health* 2012; 66(12):1182-1186.PM:22577181

50. Bellou A., Bhatt R. Reducing underage alcohol and tobacco use: evidence from the introduction of vertical identification cards. *J Health Econ* 2013; 32(2):353-366.PM:23333955

51. Bell B., Blundell R., Van Reenen J. Getting the unemployed back to work: an evaluation of the New Deal proposals. *International Tax and Public Finance* 1999; 6(3):339-360.

52. Ashenfelter O. Estimating the effect of training programs on earnings. *The Review of Economics and Statistics* 1978;47-57.

53. Heckman J.J., Ichimura H., Todd P.E. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies* 1997; 64(4):605-654.

54. Zhao M., Konishi Y., Glewwe P. Does information on health status lead to a healthier lifestyle? Evidence from China on the effect of hypertension diagnosis on food consumption. *J Health Econ* 2013; 32(2):367-385.PM:23334058

55. Berger M.L., Martin B.C., Husereaus D., Worley K., Allen J.D., Yang W. et al. A questionnaire to assess the relevance and credibility of observational studies to inform health care decision making: an ISPOR-AMCP-NCP good practice task force report. *Value in Health* 2014; 17(2):143-156.

56. Motheral B., Brooks J., Clark M.A., Crown W.H., Davey P., Hutchins D. et al. A checklist for retrospective database studies--report of the ISPOR Task Force on Retrospective Databases. *Value Health* 2003; 6(2):90-97.PM:12641858

57. Kreif N., Grieve R., Sadique M.Z. Statistical Methods For Cost-Effectiveness Analyses That Use Observational Data: A Critical Appraisal Tool And Review Of Current Practice. *Health Economics* 2013; 22(4):486-500.

58. GRACE Principles. A validated checklist for evaluating the quality of observational cohort studies for decision-making support. Accessed 1-5-2015 http://www.graceprinciples.org/doc/GRACE-Checklist-031114-v5.pdf

59. von Elm E., Altman D.G., Egger M., Pocock S.J., Getzche P.C., Vandenbroucke J.P. et al. The STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med* 2007; 147(8):594-596.

60. Griffin S.C., Claxton K.P., Palmer S.J., Sculpher M.J. Dangerous omissions: the consequences of ignoring decision uncertainty. *Health Econ* 2011; 20(2):212-224.PM:20091763

61. Abadie A., Imbens G. Large sample properties of matching estimators for average treatment effects. *Econometrica* 2006; 74:235-267.

62. Abadie A., Imbens G.W. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics* 2011; 29(1).

63. Cragg J.G., Donald S.G. Testing identifiability and specification in instrumental variable models. *Econometric Theory* 1993; 9(220):240.

64. Edidin A.A., Ong K.L., Lau E., Kurtz S.M. Mortality risk for operated and nonoperated vertebral fracture patients in the medicare population. *Journal of Bone and Mineral Research* 2011; 26(7):1617-1626.

APPENDICES

A1 NICE TECHNOLOGY APPRAISALS USING NON-RANDOMISED DATA

Table A 1: Summary of NICE Technology Appraisals using non-randomised data

Type of data	TAs	Title	Data used	Parameter	Method
IPD on intervention and comparators	TA130, TA195	Adalimumab, etanercept and infliximab (TNF-α inhibitors) for active rheumatoid arthritis.	British Society of Rheumatology Biologicals' Registry.	Transitions between states. Health utility gain from treatment.	Multivariate regression to adjust for differences in EULAR response, HAQ score at baseline, HRQoL at baseline, age, sex, disease duration, number of previous DMARDs and concomitant DMARD use.
	TA185	Trabectedin for the treatment of advanced soft tissue sarcoma	IPD on intervention: phase II trial comparing two doses of the drug. IPD on comparator: historical controls from four phase II studies.	Survival (overall, progression free and time to progression) with and without intervention	Multivariate regression to adjust for differences in the comparator arm on performance score, histopathology, age and gender.
	TA279	Percutaneous vertebroplasty and percutaneous ballon kyphoplasty for treating osteoporotic vertebral compression fractures	Medicare data.	Relative risk for death.	Propensity score matching and multivariate regression (details are not available from NICE website).
	TA304	Prostheses for total hip replacement and resurfacing arthroplasty for end stage arthritis of the hip.	National Joint Registry.	Revision rates of prostheses.	Propensity score matching on age and gender.
IPD for intervention.	TA188	Human growth hormone (somatropin) for the treatment of growth failure in children	Kabi International Growth Database	Increase in height in children with growth hormone deficiency.	Naïve comparison of means.
	TA202	Ofatumumab for the treatment of chronic lymphocytic leukemia refractory to fludarabine and alemtuzumab.	Single arm Phase II study.	Hazard ratio for overall survival and progression free survival.	Multivariate regression to adjust for age, sex, Rai score, Eastern Cooperative Oncology Group status, number of prior therapies and time since diagnosis. Treatment effect compares survival on all patients vs non responders.
	TA299	Bosutinib for previously treated chronic myeloid leukaemia	IPD from single arm Phase I/II study. Aggregate data from other studies on the comparators.	Overall survival and progression free survival.	Naïve comparison. Survival on intervention obtained from single arm Phase I/II study. Survival on comparators was informed from published studies.

Type of data	TAs	Title	Data used	Parameter	Method
	ID667	Lenalidomide in combination with dexamethasone for multiple myeloma.	Intervention arm of the RCT. Aggregate data from non-randomised study on comparator.	Hazard ratio for survival (overall, progression free survival and time to progression).	Survival estimates with intervention obtained from the intervention arm of the RCT adjusted for baseline prognostic factors. Published non-randomised studies used to obtain survival estimates for comparators. Hazard ratio for comparators calculated as the ratio in survival between comparators and intervention, adjusting for prognostic factors reported in the published studies (using mean values).
Aggregate data on intervention and comparator	TA156	Routine antenatal anti-D prophylaxis (RAADP) for women who are rhesus D negative	Meta-analysis of two published non-randomised studies.	Odds ratio for rate of sensitisation to rhesus D positive associated with RAADP	Meta-analysis used a binary logistic regression with fixed-effects model.
	TA165	Machine perfusion systems and cold static storage of kidneys from deceased donors	Published studies: one sequential cohort study and one retrospective record review	Relative risk for primary non-function, delayed graft function, graft survival and patient survival	Naïve comparison. The sequential cohort study reported graft survival rates for kidneys using the comparator during 2-years vs using the intervention in the later period. The retrospective record review reported rates using two storage solutions. Relative risks appear to have been calculated from the rates reported in the non-randomised studies.
	TA166	Cochlear implants for children and adults with severe to profound deafness	Non-randomised studies: before-and-after study and comparative non-randomised study.	HRQoL gain following cochlear implantation.	The gain in HRQoL in adults was obtained from a prospective cohort study before and after unilateral cochlear implantation. There are no details on whether any statistical techniques were employed to control for bias. The HRQoL gain in children was obtained from a survey that evaluated the parents' perception of their child's HRQoL in children with and without implants, stratified by age at implantation and duration of use of implant.

Type of data	TAs	Title	Data used	Parameter	Method
	TA209	Imatinib for the treatment of unresectable and/or metastatic gastrointestinal stromal tumours	Published studies.	Probability of death on intervention and comparator	Probability of death for interventions and comparator was obtained from individual arms of RCTs and cohort studies assuming exponential distribution.
	TA241	Dasatinib, high-dose imatinib and nilotinib for the treatment of imatinib-resistant chronic myeloid leukaemia (CML) (part review of NICE TA70), and dasatinib and nilotinib for people with CML for whom treatment with imatinib has failed because of intolerance	Published studies.	Survival (overall and progression free) on interventions and comparators	Survival on interventions and comparators was obtained from single arm Phase II trials, cohort studies and individual arms of RCTs. The methods are unclear from the FAD.
	TA242	Cetuximab, bevacizumab and panitumumab for the treatment of metastatic colorectal cancer after first-line chemotherapy	Published studies. Cohort study comparing overall survival on becazimab 1st and 2nd line vs 1st line only; cohort study comparing overall survival on bevacizumab 1st and 2nd line vs bevacizumab 2nd line; retrospective study pooling RCT and non-RCT data for cetuximab.	Survival (overall and progression free) on intervention and comparator	Non-randomised data used in mixed treatment comparison to obtain hazard ratios in the submission of one of the companies.
	TA246	Pharmalgen for the treatment of systemic reactions to bee and wasp venon allergy.	Published studies.	Risk of systemic reaction following sting.	Naïve comparison of means. RRs calculated using the rates reported in the non-randomised studies. Risk of systemic reaction following sting with the intervention is the pooled risk observed in the studies included in the SR (RCTs and non-RCTs). Risk of systemic reaction following sting without the intervention is obtained from a survey study.

IPD – Individual patient data; EULAR – European League Against Rheumatism; HAQ – Health Assessment Questionnaire; DMARD - Disease-modifying antirheumatic drugs; HRQoL – health-related quality of life.
IV – instrumental variable.

A2 COMPARISON OF CHECKLISTS TO CRITICALLY APPRAISE STUDIES USING NON-RANDOMISED DATA

Table A 2: Comparison of checklists to critically appraise studies using non-randomised data

ISPOR checklist for Retrospective Database Studies[56]	ISPOR questionnaire to assess the relevance and credibility of observational studies to inform healthcare decision making[55]	Checklist for critically appraising statistical methods to address selection bias, in estimating incremental costs, effectiveness and cost-effectiveness[57]	GRACE checklist for evaluating the quality of observational cohort studies for decision-making support[58]	STROBE checklist of items to be included in reports of observational studies in epidemiology[59]
Objective: To assist decision makers in evaluating the quality of published studies that use health-related retrospective databases. Retrospective observational studies are studies that use existing data sources in which both exposure and outcomes have already occurred.	**Objective: To create a questionnaire that would promote awareness of the issues related to alternative study designs.** Prospective observational studies are studies in which participants are not randomised or otherwise assigned to an exposure and for which the consequential outcomes of interest occur after study commencement.	**Objective: to develop a checklist to assess statistical methods for addressing selection bias in cost-effectiveness analyses that use observational data**	**Objective: To help select robust observational research of comparative effectiveness.**	**Objective: To improve the quality of reporting of observational studies and facilitate critical appraisal and interpretation.**
Questions on whether the research question answered by the study is relevant to the decision problem				
Data sources Relevance: Have the data attributes been described in sufficient detail for decision makers to determine whether there was a good rationale for using the data source, the data source's overall generalizability, and how the findings can be interpreted in the context of their own organization?	Relevance: 1. Is the population relevant?			Introduction 2. Explain the scientific background and rationale for the investigation being reported
	Relevance 2. Are any relevant interventions missing?			
	Relevance 3. Are the outcomes relevant?			
	Relevance 4. Is the context (settings and practice patterns) applicable?			
Questions on whether the data are appropriate to answer the research question proposed by the study				
Data sources Reliability and validity: Have the reliability and validity of the data been described, including any data quality checks and data cleaning procedures?	Credibility – Data 1. Were the data sources sufficient to support the study?		Data D1: Were treatment and/or important details of treatment exposure adequately recorded for the study purpose in the data sources?	Methods 5. Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection

ISPOR checklist for Retrospective Database Studies[56]	ISPOR questionnaire to assess the relevance and credibility of observational studies to inform healthcare decision making[55]	Checklist for critically appraising statistical methods to address selection bias, in estimating incremental costs, effectiveness and cost-effectiveness[57]	GRACE checklist for evaluating the quality of observational cohort studies for decision-making support[58]	STROBE checklist of items to be included in reports of observational studies in epidemiology[59]
Data sources Linkages: Have the necessary linkages among data sources and/or different care sites been carried out appropriately, taking into account differences in coding and reporting across sources?	Credibility – Data 2. Was exposure defined and measured in a valid way?		Data D2: Were the primary outcomes adequately recorded for the study purpose (e.g. available in sufficient detail through data source(s))?	Methods 6. (a) Cohort study—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up Case-control study—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls Cross-sectional study—Give the eligibility criteria, and the sources and methods of selection of participants
Data sources Eligibility: Have the authors describe the type of data used to determine member eligibility?	Credibility – Data 3. Were the primary outcomes defined and measured in a valid way?		Data D3: Was the primary clinical outcome(s) measured objectively rather than subject to clinical judgement (e.g. opinion about whether the patient's condition has improved?)	Methods 6. (b) Cohort study—For matched studies, give matching criteria and number of exposed and unexposed Case-control study—For matched studies, give matching criteria and the number of controls per case
	Credibility – Data 4. Was the follow-up time similar among comparison groups or were the differences in follow-up accounted for in the analyses?		Data D4: Were primary outcomes validated, adjudicated, or otherwise known to be valid in a similar population?	Methods 7. Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable
			Data D5: Was the primary outcome(s) measure or identified in an equivalent manner between the treatment/intervention group and the comparison group?	Methods 8. For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group
			Data D6: Were important covariates that may be known to be confounders or effect modifiers available and recorded?	
			Methods M4: Is the classification of exposed and unexposed person-time free of "immortal time bias"?	

ISPOR checklist for Retrospective Database Studies[56]	ISPOR questionnaire to assess the relevance and credibility of observational studies to inform healthcare decision making[55]	Checklist for critically appraising statistical methods to address selection bias, in estimating incremental costs, effectiveness and cost-effectiveness[57]	GRACE checklist for evaluating the quality of observational cohort studies for decision-making support[58]	STROBE checklist of items to be included in reports of observational studies in epidemiology[59]
Questions on whether the study design is appropriate for the available data and to answer the research question?				
Methods – Research design Data analysis plan: Was a data analysis plan, including study hypotheses, developed a priori?	Credibility - Design 1. Were the study hypotheses or goals pre-specified a priori?		Methods M1: Was the study (or analysis) population restricted to new initiators of treatment or those starting a new course?	Introduction 3. State specific objectives, including any pre-specified hypotheses
Methods – Research design Design selection: has the investigator provided a rationale for the particular research design?	Credibility - Design 2. If one or more comparison groups were used, were they concurrent comparators or did they justify the use of historical comparison group(s)?		Methods M2: If one or more comparison groups were used, were they concurrent comparators? If not, did the authors justify the use of historical comparisons group(s)?	Methods 4. Present key elements of study design early in the paper
Methods – Research design Research design limitations: did the author identify and address potential limitations of that design?	Credibility – Design 3. Was there evidence that a formal study protocol including an analysis plan was specified before executing the study?			Methods 9. Describe any efforts to address potential sources of bias
Methods – Research design Treatment effect: for studies that are trying to make inferences about the effects of an intervention, does the study include a comparison group and have the authors described the process for identifying the comparison group and the characteristics of the comparison group as they relate to the intervention group?	Credibility – Design 4. Were sample size and statistical power to detect differences addressed?			Methods 10. Explain how the study size was arrived at
Methods - Study Population and variable definitions Sample selection: have the inclusion and exclusion criteria and the steps used to derive the final sample from the initial population been described?	Credibility – Design 5. Was a study design used to minimize or account for confounding?			
Methods - Study Population and variable definitions Eligibility: are subjects eligible for the time period over which measurement is occurring?	Credibility – Design 6. Was the follow-up period of sufficient duration to detect differences addressed?			

ISPOR checklist for Retrospective Database Studies[56]	ISPOR questionnaire to assess the relevance and credibility of observational studies to inform healthcare decision making[55]	Checklist for critically appraising statistical methods to address selection bias, in estimating incremental costs, effectiveness and cost-effectiveness[57]	GRACE checklist for evaluating the quality of observational cohort studies for decision-making support[58]	STROBE checklist of items to be included in reports of observational studies in epidemiology[59]
Methods - Study Population and variable definitions Censoring: were inclusion/exclusion or eligibility criteria used to address censoring and was the impact on study findings discussed?	Credibility – Design 7. Were the sources, criteria and methods for selecting participants appropriate to address the study questions/hypotheses?			
Methods - Study Population and variable definitions Operational definitions: are case (subjects) and end point (outcomes) criteria explicitly defined using diagnosis, drug markers, procedure codes, and/or other criteria?	Credibility – Design 8. Were the study groups selected so that comparison groups would be sufficiently similar to each other (e.g. either by restriction or recruitment based on the same indications for treatment)?			
Methods - Study Population and variable definitions Definition validity: have the authors provided a rationale and/or supporting literature for the definitions and criteria used and were sensitivity analyses performed for definitions or criteria that are controversial, uncertain, or novel?				
Methods - Study Population and variable definitions Timing of outcome: is there a clear temporal (sequential) relationship between the exposure and outcome?				
Methods - Study Population and variable definitions Event capture: are the data, as collected, able to identify the intervention and outcomes if they actually occurred?				
Methods - Study Population and variable definitions Disease history: is there a link between the natural history of the disease being studied and the time period for analysis?				

ISPOR checklist for Retrospective Database Studies[56]	ISPOR questionnaire to assess the relevance and credibility of observational studies to inform healthcare decision making[55]	Checklist for critically appraising statistical methods to address selection bias, in estimating incremental costs, effectiveness and cost-effectiveness[57]	GRACE checklist for evaluating the quality of observational cohort studies for decision-making support[58]	STROBE checklist of items to be included in reports of observational studies in epidemiology[59]
Methods - Study Population and variable definitions Resource valuation: for studies that examine costs, have the authors defined and measured an exhaustive list of resources affected by the intervention given the perspective of the study and have resource prices been adjusted to yield a consistent valuation that reflects the opportunity cost of the resource?				
Questions on the quality of the analyses and statistical or econometric methods				
Statistics Control variables: if the goal of the study is to examine treatment effects, what methods have been used to control for other variables that may affect the outcome of interest?	Credibility – Analyses 1. Was there a thorough assessment of potential measured and unmeasured confounders?	1a. Did the study address the 'no unobserved confounding' assumption?	Methods M3: Were important covariates, confounding and effect modifying variables taken into account in the design and/or analysis?	Methods 11. Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why
Statistics Statistical model: have the authors explained the rationale for the model/statistical method used?	Credibility – Analyses 2. Were analyses of subgroups or interaction effects reported for comparison groups?	1b. Did the study assess the assumption that the instrumental variable was valid?	Methods M5: Were any meaningful analyses conducted to test key assumptions on which primary results are based?	Methods 12. (a) Describe all statistical methods, including those used to control for confounding
Statistics Influential cases: have the authors examined the sensitivity of the results to influential cases?	Credibility – Analyses 3. Were sensitivity analyses performed to assess the effect of key assumptions or definitions on outcomes?	2. Did the study assess whether the distributions of the baseline covariates overlapped between the treatment groups?		12. (b) Describe any methods used to examine subgroups and interactions
Statistics Relevant variables: have the authors identified all variables hypothesized to influence the outcome of interest and included all available variables in their model?		3. Did the study assess the specification of the regression model for costs and health outcomes?		12. (c) Explain how missing data were addressed

ISPOR checklist for Retrospective Database Studies[56]	ISPOR questionnaire to assess the relevance and credibility of observational studies to inform healthcare decision making[55]	Checklist for critically appraising statistical methods to address selection bias, in estimating incremental costs, effectiveness and cost-effectiveness[57]	GRACE checklist for evaluating the quality of observational cohort studies for decision-making support[58]	STROBE checklist of items to be included in reports of observational studies in epidemiology[59]
Statistics Testing statistical assumptions: do the authors investigate the validity of the statistical assumptions underlying their analysis?		4. Was covariate balance assessed after applying a matching method?		12. (d) Cohort study—If applicable, explain how loss to follow-up was addressed Case-control study—If applicable, explain how matching of cases and controls was addressed Cross-sectional study—If applicable, describe analytical methods taking account of sampling strategy
Statistics Multiple tests: if analyses of multiple groups are carried out, are the statistical tests adjusted to reflect this?		5. Did the study consider structural uncertainty arising from the choice or specification of the statistical method for addressing selection bias?		12. (e) Describe any sensitivity analyses
Statistics Model prediction: if the authors utilize multivariate statistical techniques in their analysis, do they discuss how well the model predicts what it is intended to predict?				
Questions on the reporting and results				
	Credibility – Reporting 1. Was the number of individuals screened or selected at each stage of defining the final sample reported?			Results 13. (a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed
	Credibility - Reporting 2. Were the descriptive statistics of the study participants adequately reported?			13. (b) Give reasons for non-participation at each stage
	Credibility - Reporting 3. Did the authors describe the key components of their statistical approaches?			13. (c) Consider use of a flow diagram
	Credibility – Reporting 4. Were confounder-adjusted estimates of treatment effects reported?			14. (a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders

ISPOR checklist for Retrospective Database Studies[56]	ISPOR questionnaire to assess the relevance and credibility of observational studies to inform healthcare decision making[55]	Checklist for critically appraising statistical methods to address selection bias, in estimating incremental costs, effectiveness and cost-effectiveness[57]	GRACE checklist for evaluating the quality of observational cohort studies for decision-making support[58]	STROBE checklist of items to be included in reports of observational studies in epidemiology[59]
	Credibility – Reporting 5. Did the authors describe the statistical uncertainty of their findings?			14. (b) Indicate number of participants with missing data for each variable of interest
	Credibility – Reporting 6. Was the extent of missing data reported?			14. (c) Cohort study—Summarise follow-up time (eg, average and total amount)
	Credibility – Reporting 7. Were absolute and relative measures of treatment effect reported?			15. Cohort study—Report numbers of outcome events or summary measures over time. Case-control study—Report numbers in each exposure category, or summary measures of exposure. Cross-sectional study—Report numbers of outcome events or summary measures
				16. a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included
				16. (b) Report category boundaries when continuous variables were categorized
				16. (c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period
				17. Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses
Questions on the interpretation of the results and discussion of the limitations, strengths and key areas of uncertainty				
Discussion/conclusions Theoretical Basis: Have the Authors Provided a Theory for the Findings and Have They Ruled out Other Plausible Alternative Explanations for the Findings?	Credibility – Interpretation 1. Were the results consistent with prior known information or if not was an adequate explanation provided?			18. Summarise key results with reference to study objectives
Discussion/conclusions Practical versus Statistical Significance: Have the Statistical Findings Been Interpreted in Terms of Their Clinical or Economic Relevance?	Credibility – Interpretation 2. Are the observed treatment effects considered clinically meaningful?			19. Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias

ISPOR checklist for Retrospective Database Studies[56]	ISPOR questionnaire to assess the relevance and credibility of observational studies to inform healthcare decision making[55]	Checklist for critically appraising statistical methods to address selection bias, in estimating incremental costs, effectiveness and cost-effectiveness[57]	GRACE checklist for evaluating the quality of observational cohort studies for decision-making support[58]	STROBE checklist of items to be included in reports of observational studies in epidemiology[59]
Discussion/conclusions Generalizability: Have the Authors Discussed the Populations and Settings to Which the Results Can Be Generalized?	Credibility – Interpretation 3. Are the conclusions supported by the data and analysis presented?			20. Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence
	Credibility – Interpretation 4. Was the effect of unmeasured confounding discussed?			21. Discuss the generalisability (external validity) of the study results
Other questions				
	Conflicts of interest 1. Were there any potential conflicts of interest?			Title and abstract 1. (a) Indicate the study's design with a commonly used term in the title or the abstract 1. (b) Provide in the abstract an informative and balanced summary of what was done and what was found
	Conflicts of interest 2. If there were potential conflicts of interest, were steps taken to address these?			22. Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based

A3 APPLICATION OF QUEENS CHECKLIST TO A NON-RANDOMISED STUDY

The QuEENS checklist is applied to a non-randomised study (Edidin *et al*[64]) used in TA279 to help inform the decision of whether percutaneous vertebroplasty or balloon kyphoplasty are effective and cost-effective interventions compared with optimal pain management for osteoporotic vertebral compression fracture (VCF). This study is one of the sources that were considered to inform the inputs used in the decision model, namely the parameter relating to the mortality effect of each of the interventions. Although no mortality effect was observed in RCTs, the company submitted three non-randomised studies that evaluated the differences in mortality between non-operated vertebral fracture patients receiving optimal pain management and patients operated (with percutaneous vertebroplasty or balloon kyphoplasty). These studies concluded that operation has a beneficial effect in mortality that has not been captured in the RCTs. Since the study used to inform the assessment group's decision model is academic-in-confidence, the only published study considered[64] is used here to exemplify the application of the checklist.

Edidin *et al* used the US Medicare inpatient and outpatient claims between 1 Jan 2005 and 31 Dec 2008 to estimate the differences in mortality rates and hazard ratios for each intervention with Cox multivariate regression over four years and differences in the probability of death over three years with IV analysis. The sample consisted of 858,978 of VCF patients, which amounted to 85.3% of the total sample of patients with vertebral fractures. A total of 182,946 (21.3%) patients were operated upon and of these 119,253 (13.9%) had a balloon kyphoplasty and 63,693 (7.4%) had a percutaneous vertebroplasty. Covariates included age, race/ethnithity; patient health status; (general - Charlson comorbidity index groups - and specific – 12 comorbidities that have been identified previously as possible causes of death associated with VCFs: arterial disease; chronic obstructive pulmonary disease; cancer; diabetes; hip fracture; hypertensive disease; ischemic heart disease; other heart disease; pneumonia; pulmonary heart disease; stroke; wrist fracture); type of diagnosed fracture (pathologic, traumatic); site of service (outpatient, inpatient); physician specialty (orthopaedic surgeon, neurosurgeon, interventional radiologist, others); socioeconomic status (per capita income for county of residence and Medicare buy-in status); year of diagnosis; and census region (Northeast, Midwest, South and West). For the IV analysis, four possible instruments were considered (physician preference, hospital preference, census region and

physician specialty) but only physician preference was found adequate and used in the analysis.

Table A 3 presents the application of QuEENS to Edidin *et al*. Q1 assesses whether different methods were compared within the study. Edidin *et al* used multivariate Cox regression and IV 2-stage regression to estimate the effect of operation, either by balloon kyphoplasty or percutaneous vertebroplasty in mortality. Q2 refers to whether the results were compared with others in the literature; estimates of the absolute mortality rates and relative risk for death were compared with other studies in the Discussion. Although the absolute mortality rates are similar with other studies using other data sources, studies evaluating the effect of the interventions on mortality did not observe a risk reduction.

Q3 assesses whether the study discussed the treatment effect identified and the assumptions required. The multivariate Cox regression assumed (i) no selection on unobservables, (ii) non-informative censoring and (iii) proportional hazards. The first assumption, that there is no selection on unobservables, means that the variables included in the regression are the only variables that affect selection into each intervention, whether to have conservative management, balloon kyphoplasty or percutaneous vertebroplasty, and outcome (mortality risk). The second assumption refers to whether censoring is related to the outcome. Censoring occurs when an individual is lost to follow-up or does not experience the event of interest within the follow-up period. Censoring is deemed to be non-informative if the individual's censoring time is independent from their outcome. Informative censoring occurs when participants are lost to follow-up due to reasons related to the study and invalidates standard survival analysis techniques such as Cox regression. The third assumption, on proportional hazards, means that the survival curves for the three strata (non-operated, operated with balloon kyphoplasty and operated with percutaneous vertebroplasty) are proportional over time. This can be evaluated graphically using 'log-log' plots. None of these three assumptions was tested or discussed.

The IV regression used physician preference as the instrument. This assumes that physician preference is correlated with treatment allocation but has no impact on outcome (mortality). However, it may be plausible that physician preference may be related with their expertise in each procedure, which in turn is related with the outcome. If physician preference is indeed related to outcome, the estimate of treatment effect may be biased. The other instruments

considered but rejected were physician specialty, census region and hospital preference. Physician specialty and hospital preference may have an impact on outcomes in a similar way as physician preference: physician specialty may be related to their expertise in a specific procedure which in turn affects outcomes whilst hospital preference may be partly determined by the hospital physicians' preferences which, as discussed earlier, may relate to their expertise. Census region was rejected because it was correlated with the outcome survival. A valid instrument should be justified both theoretically and empirically. Theoretically, in terms of the rationale behind assuming that the instrument has no effect on outcomes but is correlated with treatment assignment; and empirically, by checking the correlation of the instrument with treatment assignment. In Edidin *et al*, the instrument used was not adequately justified.

Q4 refers to the parametric assumptions of the model. Edidin *et al* used Cox regression to estimate the hazard ration on mortality associated with the interventions, which is appropriate with time-to-event analysis. The IV 2-stage regression estimated probability of death; however, no details are provided on the IV models. No checks on the model specification were reported (Q5).

The multivariate Cox regression assumes selection on observables. As discussed above, this assumption was not discussed or assessed (Q6). Minimum checks were conducted to assess the overlap of the characteristics in the treated and untreated groups. Table 2 (p.1620)[64] compares the patient characteristics but no statistical or visual checks are conducted.

Q15 to Q18 assess the application of IV methods. Q15 asks whether the instrument was well justified. Although the choice of instrument was justified (p.1623), there was no discussion on the possible influence of physician's preference on the outcome (survival). As discussed above, physician's preference is a poor instrument if it is determined by their expertise on the intervention and expertise in turn affects survival. The paper does not report statistical tests to the instrument (Q18). The sample size is sufficiently large to apply IV analysis (Q16).

Table A 3: Application of QuEENS to Edidin *et al*[64]

	Questions	Options		Comments
General issues	Q1: Have different methods been compared within the study?	(a) Yes	☑	Multivariate Cox regression and IV 2-stage regression. Multivariate Cox regression assumes no selection on unobservables whereas IV can handle selection on unobservables as long as the instrument is valid.
		(b) Partially		
		(c) No		
	Q2: Have the results of the study been compared to others in the literature?	(a) Yes, compared to alternative methods using the same dataset		
		(b) Yes, compared to similar methods using other data sources	☑	See p.1624 2nd paragraph: the absolute mortality rates observed in this study are similar with other studies using other data sources. See p.1624 3rd paragraph: other studies, using smaller datasets, did not observe a reduction in mortality as Edidin *et al* observed. The methods used by these other studies were not specified.
		(c) Not compared – no other estimates found in the literature		
		(d) Not compared		
	Q3: Is there a discussion of what treatment effect is identified and of the assumptions needed?	(a) Discussion of effect and assumptions		
		(b) Discussion of effect but not the assumptions		
		(c) Discussion of the assumptions but not the effect	☑	There is some discussion of the assumptions in the 'Discussion' (p.1619 1st and 5th paragraph and p.1624 last paragraph): in the Cox regression, patient characteristics are treated as covariates whereas IV aims to account for bias due to unobserved patient characteristics. There is discussion of the properties an instrumental variable must have in terms of their correlation with the treatment and the outcome (p1619 2nd column 1st paragraph). However, there is no discussion that the multivariate Cox regression assumes non-informative censoring and proportional hazards. The assumption that the multivariate Cox regression assumes no selection on unobservables can be implied by the discussion of the instrumental variables analysis (p.1619 1st column last paragraph until the end of the sentence).
		(d) No discussion of either		
	Q4: Is the model chosen consistent with the outcome variable if using a parametric method?	(a) Yes	☑	Multivariate Cox regression, a semi-parametric method, was used to estimate the hazard ratio on mortality due to the interventions.
		(b) Unclear	☑	The model used to estimate the probability of death

	Questions	Options		Comments
				following IV was not clearly specified.
		(c) No		
	Q5: Were any checks conducted on the model specification?	(a) Yes, appropriate (detail which)		
		(b) Yes, but inappropriate or not enough		
		(c) No checks reported	☑	No checks were reported.
Methods assuming selection on observables	Q6: On selection: Is the assumption of selection on observables assessed?	(a) Yes, expert literature or opinion cited		
		(b) Yes, theoretical reasoning given.		
		(c) No	☑	The assumption of no selection on unobservables is not appropriately discussed or assessed.
	Q7: What checks were conducted to assess overlap?	(a) Yes, thorough checks		
		(b) Yes, minimum checks	☑	Table 2 in p.1620 compares the patient characteristics; however, no statistical tests to compare differences between groups were reported. The degree of overlap between groups was not assessed.
		(c) No checks reported.		
IV methods	Q15: Is the instrument well justified? (i.e. eligibility in programme participation (reason), natural experiment, theoretically sensible, fitted propensity scores)	(a) Yes, theoretically		
		(b) Yes, citing expert literature		
		(c) No	☑	The 2nd paragraph in p.1623 explains the choice of the instrument. Four potential instruments were assessed: physician specialty, census region, hospital preference and physician preference. Physician specialty and census region were rejected because both were significantly correlated with the outcome survival; in addition, patient characteristics were unbalanced across census regions. Physician preference was preferred to hospital preference because it was more correlated with treatment than hospital preference. There was no discussion of whether the instrument physician preference could be correlated with outcome (survival).
	Q16: Is the sample size relatively large?	(a) Yes	☑	N=858,978 patients.
		(b) No		
	Q17: If more than one IV, is the test of overidentifying restrictions reported?	(a) Yes	NA	
		(b) No		
	Q18: Is a weak instrument(s) test reported?	(a) Yes		
		(b) No	☑	Tests to the potential instruments were not reported in addition to the correlation between the instrument, outcome and treatment.

Edidin *et al* found a significant reduction in mortality risk from operating people with VCF with balloon kyphoplasty or with percutaneous vertebroplasty; a higher reduction in mortality was observed for kyphoplasty. The authors hypothesised that the reduction in mortality may be related to the improvement in pulmonary function since mortality post-VCF has been attributed to pulmonary disease. Nonetheless, the mechanism of action by which operative treatment reduces mortality risk remains unclear. Another possibility is that the observed reduction in mortality may be unrelated to the intervention but determined by the patient and physician characteristics. Healthier or more active patients may be offered operative treatment whereas less healthy less active patients may be offered non-operative management. This may be related not only with expectations of better outcomes in healthier individuals but also higher risks of surgery in less healthy patients. In addition, patients with a preference for operative treatment, perhaps in the expectation of greater benefit, may chose physicians whose expertise or success rates in operative treatment are greater.

One possible method to assess whether the observed reduction in mortality risk may be attributable to treatment assignment is to compare it with the results observed in RCTs. In this case study of TA279, the evidence group identified nine RCTs, of which two were double-blinded placebo controlled. None of the studies found a statistically significant difference in mortality, which may be related to under-powering for this outcome. A meta-analysis of the three studies reporting overall mortality at 12 months found a non-significant reduction in risk (0.68 (95% confidence interval (CI) 0.30 to 1.57)). The point estimate is similar to adjusted relative risk for operative treatment vs non-operative treatment in Edidin *et al* at 0.63 (95%CI 0.62 to 0.64), which may give additional plausibility to their results.

An additional issue in the Edidin study is the counterintuitive results of the Cox regressions comparing the two (operative vs non-operative management) or three interventions. Known risk factors for mortality, such as arterial disease, diabetes, hypertension, ischaemic heart disease and stroke, appear to have a statistically significant effect in reducing mortality. This may be related to the proportional hazards assumption which may not hold. Alternatively, it may signal omitted variable bias, either from missing interaction terms or from unobserved prognostic variables that were not included in the regression.

The application of QuEENS to Edidin *et al* shed light on the key assumptions and limitations of this study to inform parameter inputs for a decision model and the decision on whether

percutaneous vertebroplasty or balloon kyphoplasty are effective and cost-effective interventions compared with optimal pain management for osteoporotic VCF. It is clear that the assumptions underpinning the methods used were not adequately justified and that the methods may not have been appropriately applied. Therefore, the estimates of treatment effect on mortality may be biased.