THE USE OF REAL WORLD DATA FOR THE ESTIMATION OF TREATMENT EFFECTS IN NICE DECISION MAKING

REPORT BY THE DECISION SUPPORT UNIT

17th June 2016

(updated 12th December)

Helen Bell[1], Allan J Wailoo[1], Monica Hernandez[1], Richard Grieve[2], Rita Faria[3]
Laura Gibson[1], Sabine Grimm[1]

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

[2] London School of Hygiene and Tropical Medicine

[3] Centre for Health Economics, University of York

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

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

ABOUT THE DECISION SUPPORT UNIT

The Decision Support Unit (DSU) 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

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

Acknowledgements

We would like to thank Pall Jonsson, Sarah Garner, Jan Phillips and others from NICE for their input to this project. Mark Pennington provided helpful comments. Warwick Evidence provided comments on the hip replacement case study, which informed this latest version. We are grateful to Jenny Dunn for formatting of the report. The authors alone are responsible for any errors or omissions.

EXECUTIVE SUMMARY

This report aims to assess the current guidance on the use of real world data (RWD) for the estimation of treatment effects in NICE decision making and identifies areas where further research or guidance is required. This report builds on the NICE Decision Support Unit (DSU) Technical Support Document (TSD17) "The use of observational data to inform estimates of treatment effectiveness in technology appraisal: methods for comparative individual patient data" (Faria *et al*, 2015), which focused on methods commonly used to estimate treatment effects from non-randomised studies, where individual patient data (IPD) is available. This report expands upon this by considering how RWD has been used to inform decision making in seven of NICE's programmes, how it could have been used and the guidance that NICE currently provides to those responsible for submitting evidence, critiquing evidence and making decisions based on those assessments.

Central to the concerns associated with the use of RWD is the issue of selection bias, which we define as bias that arises when comparing the effect of a treatment in groups that are systematically different in variables that have an independent effect on the outcome of interest (Faria *et al*, 2015). Attempts must be made at the study design, analysis and interpretation stage to try and mitigate this bias (Faria *et al*, 2015, Hernan and Robins, 2016). NICE is moving towards further use of RWD in its decision making, but selection bias is inevitable. Within the current methods guidance, there is considerable variation in the extent that different programmes deal with selection bias. However, none of the guides say how primary studies should deal with selection bias.

The three case study examples in this report highlight some of the challenges associated with minimising selection bias when using RWD to estimate treatment effects. In particular, the hip replacement example shows how attempts to minimise selection bias require early investment at the design stage in accessing large observational datasets. The MAGEC and Bosutinib examples illustrate that, if this investment is not made at the design stage, then it is very challenging to define the counterfactual so as to minimise selection bias.

The findings of this report highlight areas for future research and where further methods guidance could be developed. There should be further research into methods of synthesising

data from various sources, such as single-arm trials and historical controls, and summary statistics, and methods for handling selection bias in context of large linked, longitudinal observational datasets, and within that, methods for handling time-varying confounding and sequential treatment decisions.

CONTENTS

TABLES AND FIGURES

ABBREVIATIONS AND DEFINITIONS

DSU	Decision Support Unit
RCT	Randomised controlled trial
RWD	Real world data
MTEP	Medical Technologies Evaluation Programme
CPHE	Centre for Public Health Excellence
NCCSC	NICE collaborating Centre for Social Care
GDG	Guideline Development Group
CDF	Cancer drugs fund
EAC	External assessment centre
ERG	Evidence review group
AG	Assessment group
STA	Single Technology Appraisal
MTA	Multiple Technology Appraisal
TA	Technology Appraisal
NICE	National Institute for Health and Care Excellence
TSD	Technical support document
MAGEC	Magnetic expansion control
PROMs	Patient reported outcome measures
HES	Hospital Episodes Statistics
THR	Total hip replacement
ATE	Average treatment effect
ATT	Average treatment effect on the treated
NJR	National joint register
OLS	Ordinary least squares
BMI	Body mass index
ASA	American Society of Anesthesiologists
OHS	Oxford hip score
EOS	Early-onset scoliosis
CRG	Conventional growth rods
BBC EAC	Birmingham and Brunel Consortium External Assessment Centre
IPD	Individual patient data
HTA	Health Technology Assessment

1. INTRODUCTION

1.1. BACKGROUND AND MOTIVATION

In May 2015, the NICE Decision Support Unit (DSU) published a Technical Support Document (TSD17) "The use of observational data to inform estimates of treatment effectiveness in technology appraisal: methods for comparative individual patient data" (Faria *et al*, 2015). It provided an account of the main analytical methods available to estimate treatment effects from non-randomised individual patient level data. Whilst the report was motivated by the experiences and needs of the NICE Technology Appraisals programme, the analytical methods described and the issues for selecting and reporting those methods have applicability across NICE's work programmes. The DSU report provided recommendations aimed at improving the quality and transparency of the methods used in future appraisals or other assessments at NICE. However, the report limited itself to commonly used methods for the estimation of treatment effects from non-randomised studies that have individual patient data (IPD). This current report can be seen as an extension of the previous TSD. Here we consider the use of real world evidence for the estimation of treatment effects across NICE's programmes. Specifically, it aims to consider how "real world data" (RWD) has been used to inform decision making at NICE, how it could have been used and the guidance that NICE currently provides to those responsible for submitting evidence, critiquing that evidence and making decisions based on those assessments.

Historically, NICE, in line with many other authorities, has followed the view that randomised controlled trials (RCTs) are generally considered to provide the highest standard of evidence on comparative treatment effectiveness. However, it is not always possible or desirable to perform RCTs, for a variety of practical or ethical reasons. For example in some settings RCTs may not be able to provide estimates of treatment effectiveness on outcomes of particular interest to NICE, in populations generalizable to those for whom the technology may be used in the UK NHS, or over sufficient time scales to capture the differential impact of the interventions on relevant costs and outcomes. Even in those situations where an RCT would be feasible, it should be recognised that NICE is generally not a generator of evidence but instead reliant on the evidence that manufacturers and the research community produce. Therefore data from non-randomised sources often has to be used to estimate treatment effectiveness as part of a NICE assessment, in the absence of, or perceived insufficiency of

randomised trial data. At the heart of concerns over the use of this data is the issue of treatment selection bias due to confounding by indication (selection bias). This form of bias arises when there are systematic differences in patient or contextual characteristics between the treatments under evaluation that influence the costs and outcomes of interest. A major challenge that faces non randomised studies is that only some of these potential confounding factors are observed. A major advantage of randomisation is that, if conducted properly, observed and unobserved characteristics are balanced between the randomised groups, and so the effect of treatment versus control on the observed outcomes can be inferred. In non-randomised studies, the treatment assignment is non-random, and the mechanism of assigning patients to alternative treatments is usually unknown. Hence, in non-randomised studies the estimated effects of treatment on outcomes is subject to treatment selection bias, and this must be recognised in the interpretation of the results.

RWD is a commonly used term to describe data generated from sources that relate to everyday clinical practice, generally outside the artificial constraints of randomised controlled trials. In its broad definition, RWD can include data generated as part of pragmatic controlled trials, however most RWD does not produce randomised evidence of treatment effect. In the context of Health Technology Assessment (HTA), RWD typically presents as observational data from registries, administrative databases and surveys. Single arm clinical trials are increasingly being used to support licensing of drug therapies. These are, by definition, non-comparative, and to obtain effectiveness estimates requires a non-randomised comparison of the outcome data from the one arm trial with for example a historical cohort, and/or using a before-and-after design. Registry data may allow for a concurrent comparison of outcomes between patients receiving alternative treatments, but the potential for selection bias remains, as the individuals selected for alternative treatments of interest are liable to differ according to their prognosis.

Increasingly, RWD is being used by NICE. In some areas of Technology Appraisal the requirements by regulators for comparative, randomised data are being relaxed in order to speed up drug approval processes (Accelerated Access Review, 2016). There are also greater opportunities for RWD to supplement other types of data across NICE programmes. These new opportunities for the use of RWD bring with them the need for greater scrutiny of the underlying design and analysis of the underlying non-randomised studies to try to minimise the inevitable selection bias.

First, this report reviews the guidance that NICE issues through its "Methods Guide" manuals for Public Health Guidelines, Social Care Guidelines, Clinical Guidelines, Interventional Procedures, Medical Technologies, Diagnostics Assessment Programme and Technology Appraisal programmes. In October 2014, NICE produced a unified Guidelines methods manual which superseded the separate Public Health Guidelines, Social Care Guidelines, Clinical Guidelines and Medicines Practice Guidelines methods manuals. Other areas of guidance, such as Highly Specialised Technologies, have been excluded because they have issued very few areas of guidance, face quite different and specific methodological issues, and currently operate on interim methods. We assess the current advice on estimating treatment effects from non-randomised data, and also the most recent advice in the Public Health, Social Care and Clinical Guidelines programmes prior to the unified Guidelines manual. A summary of findings is presented in Section 0 of this report. Second, we held discussions with NICE senior management covering the programmes listed above, committee members and an external evidence review group member. A summary of the key themes that came up in our discussions are presented in Section 3 of this report. Third, we selected and describe three case studies that illustrate where observational data was used, or where its inclusion may have been warranted. The case studies are presented in Section 4 of the report. Finally, we use the previous sections of the report as evidence which informs recommendations in relation to where more specific guidance from NICE may be needed and in relation to where future methods research should be a high priority. This is presented in Section 0.

2. REVIEW OF NICE METHODS MANUALS

2.1. INTRODUCTION

The purpose of this section is to review the guidance provided by NICE via its methods manuals. We summarise and assess the guidance provided on the use of real world (non-randomised) data for the following guidance producing programmes;

- Public Health
- Social Care
- Clinical Guidelines
- Interventional Procedures
- Diagnostics Assessment Programme
- Medical Technologies Evaluation Programme (MTEP)
- Technology Appraisals

It is important to note that NICE have replaced the separate methods manuals for Public Health, Social Care and Clinical Guidelines with a unified Guidelines methods manual. Our review will cover the unified Guidelines manual and the most recent Public Health, Social Care and Clinical Guidelines methods manuals prior to the unification.

First, we summarise each programme in terms of its remit, process and relevant methods guidance. Second, we compare the methods guidance across and within the programmes, highlight areas where they differ and establish if this can be justified based on the characteristics of the programme. Finally, we suggest areas where changes may be warranted. A summary of key differences between the methods manuals is provided in Table 1.

Table 1: Summary of references to non-randomised evidence and assessment of bias in NICE methods manuals

Characteristics of the methods guidance	Former Manuals			Current Manuals				
	Public Health[(i)]	Social Care[(ii)]	Clinical Guidelines[(iii)]	Unified Guidelines[(iv)]	Interventional Procedures[(v)]	Diagnostics Assessment[(vi)]	MTEP[(vii)]	Technology Appraisals[(viii)]
States a preference for RCTs, when appropriate RCT evidence is available	✓	✓	✓	✓	✓	✓		✓
States that non-randomised evidence could be considered	✓	✓	✓	✓	✓	✓	✓	✓
Implies that non-randomised studies should be included routinely in a systematic review	✓	✓	✓	✓	✓	✓	✓	
Identifies types of non-randomised study that are common in the programme	✓	✓	✓	✓	✓	✓		
Definition of non-randomised study types	✓			✓				
States that observational or non-randomised studies may be biased	✓	✓	✓	✓	✓	✓		✓
Defines the types of bias that may appear in non-randomised studies						✓		
Includes a checklist to assess the quality of non-randomised studies	✓	✓	✓					
References other sources of guidance for assessing the quality of non-randomised studies	✓		✓	✓		✓		
Provides guidance on how to reduce bias in non-randomised studies for estimating treatment effects								

[(i)]Methods for the development of NICE Public Health Guidance (third edition).[(ii)]The social care guidance manual. [(iii)]The guidelines manual. [(iv)]Developing NICE guidelines: the manual. [(v)]Interventional Procedures programme manual.[(vi)]Diagnostics Assessment Programme manual. [(vii)]Medical Technologies Evaluation Programme: Methods guide. [(viii)]Guide to the methods of technology appraisal 2013, Guide to the processes of technology appraisal.

2.2. FORMER MANUALS

In the following subsections, we describe the most recent Public Health, Social Care and Clinical Guidelines methods manuals prior to the publication of the unified Guidelines manual.

2.2.1. Public Health

This subsection describes the guidance provided in the "Methods for the development of NICE public health guidance (third edition)," published in September 2012. This manual has now been superseded by the unified Guidelines manual, "Developing NICE guidelines: the manual," published in October 2014. The Public Health programme makes recommendations

on what is known from research and practice about the effectiveness and cost effectiveness of interventions that adjust human behaviours to reduce the risk of potentially preventable diseases. Potential topics for appraisal are proposed by stakeholders and NICE's internal topic selection group decide which topics are carried forward to appraisal. Evidence is collated by the Centre for Public Health Excellence (CPHE) evidence review group and presented to one of their standing committees for appraisal.

The evidence presented to the committee is derived from a number of sources and covers an array of pre-specified research questions. Table 2 shows the type of evidence that should be used to address different types of research question and the type of standard CPHE review it will involve. Each of the identified research question themes can rely on a review of observational studies to address the question. According to the former guidance, an observational study review can either be an epidemiological review, a correlates review, or an effectiveness review. The effectiveness review can include evidence from observational studies, qualitative studies, and experimental studies.

Table 2: Type of evidence in Public Health appraisals

		Type of evidence					
		Systematic review of effectiveness (and cost effectiveness)	Experimental study	Observational study	Qualitative study	Practice or case report	Economic or cost-effectiveness study
Research question	Extent of public health problem or issue			CPHE epidemiological review			
	Factors or determinants or associations	CPHE review of reviews		CPHE correlates review			
	Intervention effectiveness or cost effectiveness	CPHE review of reviews	CPHE effectiveness review (can include observational and qualitative studies as well as experimental studies)				CPHE cost-effectiveness review
	Views and experiences of practitioners			CPHE correlates review	CPHE review of qualitative evidence	CPHE mapping report	
	Views and experiences of target population			CPHE correlates review	CPHE qualitative review		

The Public Health guidance stated that "the randomised controlled trial (RCT) is normally the most appropriate source of evidence for judging the 'efficacy' of clearly circumscribed interventions that are implemented in ideal circumstances. However, such evidence is not

always available or appropriate: it may not be feasible to conduct RCTs for some complex, large-scale, multi-agency and multi-faceted interventions, policies and services; and in some cases it may be unethical to do so. Further, given the complexity of causal chains in public health, the external validity of some RCT findings often has to be enhanced by observational studies to determine the 'effectiveness' of interventions in real-life situations. For evaluating large-scale interventions, observational studies may be the only feasible option (Victora *et al*. 2004).” (Methods for the development of NICE Public Health Guidance third edition, Pg 18). Hence, we can conclude that observational data is a commonly presented form of evidence in Public Health appraisals. The manual refered the reader to Medical Research Council (MRC) guidance on evaluating complex interventions (Craig *et al*. 2008) and using natural experiments to evaluate population health interventions (Craig *et al*. 2011).

The evidence is generally taken from published studies, however evidence from unpublished studies, research in progress and the grey literature is sometimes considered. The commonly used types of non- randomised quantitative studies were listed as;

- Before-and-after study.
- Non-randomised controlled trial (NRCT).
- Before-and-after study.
- Case–control study.
- Cohort study.
- Correlation study.
- Cross-sectional study.
- Interrupted time series.

The reviewers are required to assess the quality of evidence. The guidance provides a section on `Assessing the quality of evidence’, a flow chart to identify the type of study (see appendix 1), and a checklist with guidance to appraise the quality of quantitative studies (see Methods for the development of NICE public health guidance (third edition), Appendix F). However, there is no guidance is provided on how the primary study should be designed and analysed to minimize selection bias from non-randomised data.

2.2.2. Social Care

This subsection describes the guidance provided in the "The social care guidance manual," published in April 2013. This manual has now been superseded by the unified Guidelines manual, "Developing NICE guidelines: the manual," published in October 2014. Social care generally refers to all forms of personal care and other practical assistance for children, young people and adults who need extra support (The social care guidance manual, pg10). Topics for appraisal are referred from the Department of Health or the Department of Education. The scope is defined by NICE Collaborating Centre for Social Care (NCCSC) representatives, the Guidance Development Group (GDG) chair (once appointed), the GDG Social Care topic adviser (if there is one) and NICE representatives. Information specialists from the NCCSC conduct a systematic review to obtain evidence from published and unpublished studies, studies in progress, conference abstracts, legislation and grey literature. Evidence can also be submitted by registered stakeholders. The evidence is presented to the GDG and used to inform the development of guidance.

The former Social Care guidance manual recognised that in the context of social care, RCT evidence needs to be supplemented by evidence from other sources. It suggested that NICE may consider evidence from qualitative studies, practitioner views and experiences to assess factors that may affect the real world applicability of the intervention. Clinical and epidemiological evidence may be used to examine outcomes, context, process and adoption (implementation), as well as barriers to and facilitators of interventions. The manual acknowledged that "there is little academic consensus about how best to synthesise information from different study designs or research models or about how to use the evidence synthesis to develop guidance."

The manual provided checklists for assessing cohort studies and case-control studies (see The social care guidance manual, Appendices D and E). The case-control study checklists were almost identical to those used in the Public Health programme, whereas the cohort study checklist had a greater focus on detecting bias. The types of bias referred to are: selection bias, attrition bias, performance bias and detection (non-blinding of assessors, etc.) bias. The checklist guidance provided an outline of the factors that may indicate the presence of bias, but assumed that the reader of the checklist has a good knowledge of the methods for assessing non-randomised data.

2.2.3. Clinical Guidelines

This subsection describes the guidance provided in the "The guidelines manual," published in November 2012. This manual has now been superseded by the unified Guidelines manual, "Developing NICE guidelines: the manual," published in October 2014. The Clinical Guidelines programme focuses on the management of a particular disease or condition. They make recommendations on care that is most suitable for the majority of patients with the disease or condition. Service guidance is also covered under the remit of the Clinical Guidelines programme, which focuses on the provision and configuration of clinical services. The advice covers a broad range of factors, including treatments, technologies and lifestyle advice. Hence, the recommendations made by Clinical Guidelines may be related to advice provided in other NICE programmes. The coordination of Clinical Guideline development is either undertaken within the Centre for Clinical Practice (CCP) at NICE or commissioned to a National Collaborating Centre (NCC). The CCP/NCC are responsible for reviewing the evidence and assessing the quality. A Guideline Development Group (GDG) contributes to preparing the scope, assesses the evidence presented by the CCP/NCC and develops the guidance. Stakeholders, patients and public are consulted during the guidance development.

The manual stated "A review question relating to an intervention is usually best answered by a randomised controlled trial (RCT), because this is most likely to give an unbiased estimate of the effects of an intervention…. There are, however, circumstances in which an RCT is not necessary to confirm the effectiveness of a treatment because we are sufficiently certain from non-randomised evidence that an important effect exists. This is the case only if all of the following criteria are fulfilled:

- An adverse outcome is likely if the person is not treated (evidence from, for example, studies of the natural history of a condition).
- The treatment gives a dramatic benefit that is large enough to be unlikely to be a result of bias (evidence from, for example, historically controlled studies).
- The side effects of the treatment are acceptable (evidence from, for example, case series).
- There is no alternative treatment.
- There is a convincing pathophysiological basis for treatment." Pg 60-61

Although the manual provided a list of situations where non-randomised data can be accepted, there is no mention here that non-randomised studies should be designed to minimise bias. The omission of this criterion implied that poorly designed non-randomised studies could potentially be considered sufficient if they meet the other criteria. The updated unified Guidelines provides more context with some specific examples of cases that may meet the criteria.

The manual highlighted different types of non-randomised study that may be assessed in the programme, including cross-sectional studies, cohort studies, case-control studies and case-series. In line with the former Social Care and Public Health methods manuals, the former Clinical Guidelines manual also included checklists for assessing the quality of cohort studies and case-control studies in an appendix. These checklists duplicated the Social Care checklists. The Clinical Guidelines manual included a reference to the quality appraisal criteria from Quality Assessment of Diagnostic Accuracy Studies (QUADAS), which is also referenced in the Diagnostics Assessment Programme methods manual. The manual advocated that the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach should be used to assess the quality of clinical effectiveness evidence and interventional studies. "The Cochrane handbook for systematic reviews of interventions" (Higgins and Green, 2011) was referenced in the guidance. The handbook includes a chapter on reviewing non-randomised studies, although this section is not specifically referred to.

2.3. CURRENT MANUALS

In the following subsections, we describe the methods manuals that were current at the time that this report was published.

2.3.1. Unified Guidelines

The unified NICE Guidelines manual provides advice on making evidence based recommendations across five of NICE's programmes, including Public Health, Social Care, Clinical Guidelines, and Medicines Practice. This methods manual supersedes the Public Health, Social Care, Clinical Guidelines manuals described in subsections 0, 2.2.2 and 2.3. The manual states that Guidelines committees are either formed from members of NICE standing Committees, with additional recruitment of topic expert members, or as topic-specific Committees with multidisciplinary recruitment. The Evidence Review Group (ERG) collate evidence based on a predefined and approved review protocol, which stipulates the

inclusion and exclusion criteria, data extraction methods, quality assessment and strategies for data synthesis. Registered stakeholders may also submit evidence for the consideration of the committee. The manual advises that a meta-analysis is performed by the evidence review group when treatment effects from more than one study are available, and network-meta-analyses are performed when only indirect comparisons are available.

The quality of the evidence must be rated by the ERG. In relation to quantitative evidence, the manual refers the reader to "The Cochrane handbook for systematic reviews of interventions" (Higgins and Green, 2011). A series of website links in appendix H of the Guidelines methods manual provide further guidance on the assessment of quantitative evidence. The references include the quality appraisal criteria from QUADAS and the GRADE approach. Both were referenced in the former Clinical Guidelines manual and QUADAS is also referenced in the Diagnostics Assessment Programme methods manual. The unified Guidelines manual refers to assessing the internal and external validity of the studies and mentions that the risk of bias should be assessed, but unlike the former Public Health, Social Care and Clinical Guidelines method manuals, the unified Guidelines manual does not provide checklists to assess the evidence (although it does refer the reader to other sources). Furthermore, there is no distinction made between the specific types of biases that may occur, unlike the former Social Care method manual.

2.3.2. Interventional Procedures

An interventional procedure involves an incision, a puncture or entry into a body cavity, or use of ionising, electromagnetic or acoustic energy. The Interventional Procedures programme aims to assess the safety and efficacy of interventional procedures that are not currently standard clinical practice. The programmes technical team perform an initial screening of the evidence to select a small number of key studies to present to the committee. The guidance states that safety and efficacy evidence from RCTs is preferred, however evidence from non-randomised studies, studies without a comparator (e.g. case series or case report) or registry data may be considered. Registry data must meet the standards outlined in

Table 3.

The guidance highlights the following areas for consideration during the evidence review:

“ • patient selection
- patient enrolment or recruitment method (for example, whether it was continuous)
- previous operator training for the procedure
- previous volume of experience of operators or participating units with the procedure
- relevance of outcomes measured
- validity and reproducibility of measurement of outcomes (for example, blinding)
- appropriateness of analysis (for example, intention-to-treat analysis)
- completeness of follow-up, for any studies involving post-procedure follow-up
- reasons for loss to follow-up
- general considerations about validity and generalisability of the studies
- inclusion of the same patients in more than 1 study
- multiple reporting of a single study
- other potential sources of bias.” (Interventional Procedures programme manual, Pg 38)

The criteria for assessing the quality of the studies is far less detailed and more subjective than in Public Health, Social Care, Clinical Guidelines and the unified Guidelines manuals. There is no specific guidance on assessing bias. The Interventional Procedures guidance states that “while several critical appraisal checklists exist, it is difficult to be prescriptive about using such lists because the relative importance of the issues varies according to the procedure, the indication and the available evidence.” (Interventional Procedures programme manual, Pg 38)

The manual also states that “sometimes, all the evidence for a procedure is from non-comparative studies (for example, reports of case series). Selected evidence about key efficacy and safety outcomes of established practice may then be presented.” (Interventional Procedures programme manual, Pg 34). However there is nothing to suggest how to assess bias and ensure that the patients receiving the intervention versus established practice are similar according to prognostic characteristics.

When further evidence about safety and efficacy is required, NICE may request that additional evidence is collected in formal clinical studies or through routine data collection through a register. Again, registry standards outlined in

Table 3 must be met. There is no advice on how safety and efficacy should be assessed from these observational data.

Table 3: Registry standards and criteria for recommending a register in Interventional

Standards	Criteria
All known procedures (all devices), without exception, are recorded in the database	Raw anonymised data available for secondary analysis and validation. Denominator data available to assess data coverage, such as sales figures and routine health service information.
The data recorded address relevant efficacy and safety outcomes and important patient characteristics	Medicines and Healthcare products Regulatory Agency/NICE and professional representatives involved in dataset design and agree final protocol. Data include details of modifications or evolution of procedure/device and numbers done for the original indication (and respective outcomes).
Independent oversight	Independent steering group responsible for design, data monitoring and analysis. Register recorded on national database of registers. Explicit intent to publish results whatever the outcome. Process for data collection, storage and analysis independent of any particular company or any commercial interest.
The Register must comply with the data protection principles laid out in the UK Data Protection Act 1998 and any other relevant legislation	Data is: used fairly and lawfully used for limited, specifically stated purposes used in a way that is adequate, relevant and not excessive accurate kept for no longer than is absolutely necessary handled according to people's data protection rights kept safe and secure not transferred outside the European Economic Area without adequate protection.

2.3.3. Diagnostics Assessment Programme

The Diagnostics Assessment Programme covers recommendations on the use of diagnostic technologies, which include tests and monitors. Evidence on technology accuracy, technology side effects and, effectiveness of treatment pathways can be considered. The guidance states that studies which test accuracy are generally cohort, cross-sectional or retrospective case-control studies. "Most compare a single index test of interest with a reference standard in order to calculate the accuracy. Paired design studies compare two index tests with each other, and often also with a reference standard. These studies are less prone to bias resulting from confounding but are rarely available." (Diagnostics Assessment Programme manual, Pg 71)

Technology side effects may be identified from RCTs, other comparative studies, cross-sectional studies, case studies and patient registries. The guidance states (pg71) that treatment effectiveness evidence from RCTs or meta-analysis of multiple RCTs is preferred to other

comparative designs, such as controlled studies, cohort studies and case-control studies, because these types of study are at risk of bias. The manual highlights and defines a number of different biases some of which are particularly relevant to diagnostic technologies.

The manual notes that each study included in the systematic review should be critically appraised to assess the validity of its results. The reader is referred to the QUADAS checklist for critically appraising diagnostic accuracy studies.

There is an integrated research facilitation process covering the Diagnostics Assessment Programme and MTEP. This allows the facilitation of research to address specific uncertainties important to the committee, conducted in a time scale of three years from research recommendations to guidance review date, with an expectation that the research generated will inform the guidance review process. However, the problem of selection bias should be carefully considered when designing these additional studies.

2.3.4. Medical Technologies Evaluation Programme (MTEP)

MTEP is responsible for making recommendations on the use of medical technologies. Manufacturers (or sponsors) notify NICE of their product and eligible technologies are routed to the appropriate appraisal programme. The External Assessment Centre (EAC) critiques evidence submitted by the sponsor, may assess additional evidence and/or perform additional analyses. Quantitative evidence can include published studies and unpublished evidence sources (such as "observational research sources, including professional or manufacturer sponsored registers" (Medical Technologies Evaluation Programme: Methods guide, Pg 14)).

Medical Technologies differ from other medical interventions in a number of ways. They may be modified in ways that change their effectiveness and there may be a learning curve in terms of the user gaining competence in the use of the technology over time. Medical Technologies are rarely evaluated in RCTs because there is no requirement for such evidence for CE marking. Many companies are small and do not have the resources to conduct such studies. Hence there are often situations when evidence from multiple sources needs to be combined to assess the effectiveness of a technology relative to a suitable comparator. MTEP has no threshold on the quality or design of studies submitted as evidence. They may be published or unpublished. However, there is no specific advice on how effectiveness should

be assessed when the evidence on intervention and evidence on the comparator come from different sources.

The methods described in the MTEP manual are less defined than some of the other programmes. There is no specific advice on how published and unpublished evidence from existing studies should be assessed for bias. Furthermore, there is no specific mention of how observational data from registries may be appropriately used to provide evidence.

In common with the Diagnostics Assessment Programme, the MTEP programme is also covered by the integrated research facilitation process. Again, the problem of selection bias should be carefully considered when designing these additional studies.

2.3.5. Technology Appraisals

The Technology Appraisal programme makes recommendations on the use of new and existing medicines and treatments within the NHS. Technologies can be assessed as Single Technology Appraisals (STAs) or Multiple Technology Appraisals (MTAs). In STAs, manufacturers submit evidence which is reviewed by an evidence review group (ERG). In MTAs, the evidence is collated and assessed by the assessment group (AG). In Technology Appraisals, evidence from randomised controlled trials (RCTs) is considered to be the most appropriate measure of a relative treatment effect. However, the manual states that RCTs may not always provide sufficient evidence to quantify the treatment effect, therefore data from non-randomised studies may be required to supplement RCT data. The guidance also states that a systematic review should include all relevant evidence available, including evidence from non-randomised studies, and evidence from the included studies can be synthesised in a meta-analysis.

The Technology Appraisal methods guide provides the following statement on non-randomised data;

“The problems of confounding, lack of blinding, incomplete follow-up and lack of a clear denominator and end point occur more commonly in non-randomised studies and non-controlled trials than in RCTs.

Observational (or epidemiological) studies do not apply an intervention, but instead compare outcomes for people who use the technology under appraisal with outcomes for people who do not use the technology. These studies may be biased in that the people who use the technology may fundamentally differ in their risk of the outcome than the people who do not use the technology. Some observational studies lack a control group, and include only people who receive the technology.

Inferences will necessarily be more circumspect about relative treatment effects drawn from studies without randomisation or control than those from RCTs. The potential biases of observational studies should be identified, and ideally quantified and adjusted for. When possible, more than 1 independent source of such evidence should be examined to gain some insight into the validity of any conclusions. …Study quality can vary, and so systematic review methods, critical appraisal and sensitivity analyses are as important for review of these data as they are for reviews of data on relative treatment effects from RCTs."

(Guide to the methods of technology appraisal 2013, Pg 22-23)

It is clearly stated that caution is necessary when assessing evidence from observational studies due to biases, however there is no specific guidance on how to assess the studies for bias. The methods guidance does not describe alternative approaches to the design and analysis of observational studies that can be used to estimate treatment effects, nor how analysts should try and minimise the inevitable role of selection bias in the design, analysis and interpretation of results.

2.4. DISCUSSION

This review of the NICE methods manuals illustrates the extent of variation in guidance for assessing evidence from non-randomised data sources. All programmes except MTEP specifically state a preference for RCTs, however it is generally accepted across the programmes that systematic reviews should include evidence from non-randomised studies. The Public Health and Guidelines manuals define types of non-randomised study design and the Diagnostics Assessment Programme defines types of biases that are likely to appear. The former Public health, Social Care and Clinical Guidelines manuals provided checklists for assessing the quality of existing non-randomised studies, however there was some potential for improvement in the checklists. In order to avoid duplicating work done externally, the unified Guidelines manual, which superseded the Public health, Social Care and Clinical

Guidelines manuals, does not include multiple checklists but rather refers to external sources for appropriate and up-to-date checklists. None of the manuals provides guidance on how evidence from an intervention should be compared with evidence from a comparator when there is only single-arm/case series evidence available on the intervention. Also, none of the methods guides provides specific advice on how the primary study should be designed, analysed, and interpreted to minimise the selection bias due to confounding, that is inevitable with non-randomised studies.

The variation in the extent of guidance on non-randomised data is partly due to the remits of the different programmes and frequency that non-randomised data is used. Public Health and Social Care tend to assess large complex interventions and therefore are highly reliant on observational studies. RCT evidence for the relevant populations of these interventions is often unavailable, not feasible or not desirable for a variety of reasons. As a consequence their guidance was more equipped to deal with the assessment of observational studies, and the majority of this detail is present in the unified Guidelines manual, which supersedes the separate manuals for Public Health and Social Care. When RCT evidence is available, it is likely to depict a subset of the population, therefore supplementary evidence is required from observational studies to understand how the evidence would translate to the real-world scenario.

Technology Appraisal in practice adopts a more rigorous evidence hierarchy because clinical trials are linked to the product license in most situations and therefore tend to be available to inform a substantial element of the evidence submission. However, this does not ensure that trial populations are always reflective of relevant NHS patient populations or the subgroups of patients from within the licensed indication.

The evidence submitted to the diagnostic, Interventional Procedures and Medical Technologies Evaluation Programmes tends to involve more instances of non-randomised evidence. This is due to the nature of the products evaluated in these programmes as RCT evidence is not always available. Sometimes the only available evidence is case series with no comparator, yet there is no specific methods guidance in any of these programmes regarding how bias can be minimised when establishing a suitable comparator. A simple comparison of two means is unlikely to be sufficient.

It is generally accepted across the programmes, that systematic reviews may include non-randomised studies. However, there is also an indication that, for the estimation of treatment effects, it may be sufficient to restrict attention only to RCT data (as is the case for example, in the Technology Appraisal Methods Guide). Not all programmes attempt to explain how bias in existing studies should be assessed. The Technology Appraisal manual states that observational studies are subject to potential bias, without including specific detail on how bias can be mitigated and assessed. This statement alone, without sufficient context, may influence reviewers and committee members' willingness to accept data from non-randomised studies. In contrast, the former Social Care and Public Health guidance manuals go much further than the other programmes, by providing checklists to assess the quality of the evidence and identify bias. The checklists provide a series of questions to assess studies against, some of which relate to potential biases. The checklists are not included in the unified Guidelines manual. Other programmes, and those covered by the unified manual, could benefit by including similar checklists in their methods manuals.

In addition, the manuals across the programmes could benefit from alignment of their descriptions of common study types and associated sources of bias. Although the differences in the extent of guidance partly stems from differences in remit and frequency that non-randomised data is used in each programme, this does not act as justification for these differences to persist. Rational decision making requires consistency and resources are being wasted if poorly designed studies are tolerated within one programme but not within another. The unified Guidelines manual provides a step towards this aim.

3. INFORMAL INTERVIEWS

In order to assess the use of real world data in NICE decision making more thoroughly, we held a series of informal interviews with a small number of NICE senior management and committee members. Numbers of interviews were constrained by the timelines available for this project. We sought views on current practice in the use of real world data and areas of recurrent concern. We also asked about future challenges; how these can be dealt with and what needs to be prepared, in terms of methods guides and associated NICE support, in response to these. We also sought views on topics that could serve as suitable case studies.

We held discussions with five people: two representing senior management at NICE covering Social Care, Public Health, Medical Technologies, Interventional Procedures and Diagnostics Assessment. We spoke to a member of each of the Public Health Guidelines Committee and the Technology Appraisal committee. We also spoke to one member of an academic group responsible for critiquing evidence for several of NICE's programmes.

3.1. ISSUES HIGHLIGHTED BY PARTICIPANTS

During the conversations with the interview participants, a number of issues were discussed. These issues have been grouped into the following categories; situations when RWD is used, types of RWD used, future challenges and other issues. These are discussed in the remainder of this sub-section.

3.1.1. Situations when RWD is used

Although RCTs are generally considered to be the best form of evidence, there are often circumstances within the NICE programmes where alternative evidence needs to be considered. In the Public Health, Social Care, Interventional Procedures, Diagnostics Assessment and Medical Technologies programmes, it is common for RCT evidence on a treatment effect of an intervention relative to a suitable comparator to be unavailable or insufficient. In the Technology Appraisal programme, RCT evidence is relied upon in the vast majority of cases, however it is sometimes necessary to use real world data to provide further support for the RCT evidence, and in rare cases where RCT data is not available. The following subsections summarise the main reasons that RCT may be insufficient for each of the programmes based on our discussions.

3.1.1.1. Public Health and Social Care

- In the Public Health and Social Care programmes, it is not always possible to randomly expose individuals/patients to the intervention since all individuals may already be exposed to the intervention.
- Randomised controlled trials were perceived more difficult to finance.
- When Public Health and Social Care RCTs do exist, they are often small and the intervention is applied to specific subsets of the population, and therefore may not be applicable to the entire target population.

3.1.1.2. Medical Technologies, Interventional Procedures, Diagnostics Assessment

- The interviewees mentioned that these programmes rarely have access to RCT evidence due to the nature of the technology. The companies that make submissions to MTEP, Interventional Procedures and the Diagnostics Assessment Programme tend to be small and often cannot afford to fund RCTs. There are often lower entry costs involved in setting up these types of companies, compared to pharmaceutical companies, and there was the concern that restricting the evaluation to RCT evidence would lead to bias against effective technologies.
- Evaluations in these programmes can be affected by learning curve effects. The learning curve refers to improvement in the operator's ability to use a technology or diagnostic system, or carry out a procedure, as they become more familiar with doing so. Confounding in a trial can be caused if the learning curve is still operating during the trial period, therefore the benefit of novel technology, procedure or diagnostic system may be underestimated.
- There is not always a suitable comparator in current use. In these cases, it can be difficult to develop placebo devices or unethical to use a placebo device or procedure.

3.1.1.3. Technology Appraisals

- RCT evidence is used for the overwhelming majority of relevant outcomes. One disease area where observational data is routinely used in drug appraisals is for Hepatitis C. In this case, RCTs were seen by interviewees as not appropriate as there is a low cure rate without treatment, and low adherence rates of standard care treatments due to side-effects.
- A similar situation may occur in some immuno-oncologic treatments, where the expectation of benefit is very high compared to the current standard treatment. In situations where treatment has a direct impact on survival and clinical equipoise is lost, it is not considered to be ethical to randomise patients to a less beneficial treatment. Hence, randomised trials may be confounded by high proportions of treatment switching and single-arm trials are more common in immuno-oncologic treatments than in some other treatment areas.
- The interviewee held the view that all other disease areas generally require RCT evidence. The companies that submit the evidence may highlight some real world

observational data that indicates that the treatment effect estimated for the RCT underestimates the treatment effect when applied to the population.

- The interviewee mentioned the following reasons for supplementing RCT evidence with evidence from the real world for the following reasons;
    - Adherence rates – patients in trials are more likely to take all of the medicine prescribed, whereas patients in the wider population may adhere to it less. The reduced dose in the control group may overestimate the effectiveness of the control group (or experimental) treatment, hence if the overestimation is disproportionate in each group (i.e. if the adherence rate is likely to differ in the control group and treatment group to the general population by differing amounts, maybe due to the way that the treatment is administered) the RCT treatment effect may not reflect the true treatment effect when applied to the entire population.
    - The eligibility criteria of trials may mean that a trial only represents a subset of the population of interest.
    - Exposure to lines of previous therapy may have an impact on the performance of a treatment.
- Extension studies – Once an RCT is completed, data may continue to be collected on the experimental arm of the trial to test for maintenance of effect, which is relevant to chronic diseases. There is risk of bias if attrition or discontinuation differs by prognostic factors.

3.1.2. Type of RWD used

The discussions emphasised the distinct differences in the type of evidence assessed by each of the programmes. Some of the programmes source the majority of their RWD evidence from the published literature, whereas other programmes have directly used observational data to estimate treatment effects, to provide evidence of comparators or support RCT evidence. This sub-section provides an overview of the type of data that has been used in the programmes;

3.1.2.1. Public Health and Social Care

Observational evidence in these programmes largely comes from systematic reviews of published and unpublished studies, however evidence from other sources may also be

considered if it is relevant. Raw observational data, from registries or other sources, is not generally analysed by the review group. The evidence is assessed by the review group according to the Cochrane Handbook (Higgins and Green 2011) section on non-randomised studies and the quality assessment checklists included in the NICE methods manuals.

Although there is often a large amount of epidemiology and population evidence provided to give context to the intervention, the actual treatment effect used in the cost effectiveness analysis often relies on one study or a meta-analysis of multiple studies. Sometimes meta-analyses are performed by the review group when there is no published meta-analysis available. It is important that sensitivity analyses are undertaken to assess the uncertainty around the estimates, however this is sometimes overlooked.

3.1.2.2. Medical Technologies, Interventional Procedures, Diagnostics Assessment

The majority of observational evidence in these programmes comes from published studies that have performed an analysis comparing an intervention with a comparator on observational data. However, it may not always be the case that the statistical analysis has been performed in a way that minimises bias. Although the review teams critique the studies, they do not usually critique the statistical methods. The NICE methods manuals for these programmes do not provide specific guidance on assessing the quality of observational studies. Due to the nature of these programmes, the evidence on the intervention is often presented as single arm case-series studies and is often poor quality. Evidence for comparators is often taken from meta-analyses of comparator studies. When the evidence presented is questionable, the external assessment group can be asked to do further statistical analyses. However, this is often limited to meta-analyses. In-house analysis of registry data is costly and time consuming.

3.1.2.3. Technology Appraisals

In the Technology Appraisals programme, non-RCT data is rarely used to estimate treatment effects, and evidence from observational studies is not usually included in systematic reviews. The main purpose of observational data within the programme is to extrapolate RCT data to the longer term, such as whether a treatment effect is sustained over a longer follow-up period. The data is generally aggregated and does not provide detail about which types of patients maintain an effect. Observational data from Hospital Episode Statistics (HES) or registries may be used to provide evidence of control group relapse rates for example. The

use of non-RCT efficacy data or other clinical evidence is most common for devices (e.g. insulin pumps, cochlear implants, endovascular stents), interventions where RCTs are difficult (such as Anti-D treatments or venom prophylaxis), and in conditions with poor prognosis, where single arm studies are often carried out (e.g. sarcomas, GIST, resistant leukaemias). Novel treatments have been compared to historical data to estimate cure rates in hepatitis C drug appraisals.

3.1.3. Future challenges

NICE are seeking to make decisions earlier in the drug development path, speeding the possibility of routine NHS use for new products and interventions. But with earlier assessment comes greater uncertainty, therefore managed entry agreements (commissioning through evaluation, post-market surveillance, price agreements) have been proposed in a number of programmes, including the Interventional Procedures, devices and Technology Appraisals. The new arrangements by which drugs may be used through the Cancer Drugs Fund are a clear example of a managed entry agreement with additional data collection required (which may be real world or additional RCT evidence, in principle).

Managed entry agreements that entail evidence collection refer to a situation where NICE agrees to temporarily fund a treatment, device or procedure over a pre-specified term (e.g. 2 years). During this period, data on the treated patients are collected and submitted to a database for analysis. At the end of the period, the data is analysed and NICE decide whether or not to continue to recommend the treatment based on cost effectiveness. In the future, there is likely to be more managed entry agreements for assessing the benefit of the intervention. Hence, the challenge is to address the practical issues surrounding the collection and analysis of data, so as to minimise the inevitable selection bias and provide accurate estimates of relative effectiveness and cost-effectiveness.

3.1.4. Other issues

The assessment of the quality of evidence presented to the committee often relies heavily on the assessment made by the academic review group for the programme. The discussions revealed that the statistical methods used to assess observational studies were not critiqued in detail by all review groups. Although the reviewers are often skilled in systematic review and meta-analysis using RCT data, they may lack the specialist expertise to critically assess the quality of observational studies or to perform a statistical analysis on patient-level

observational data. Some of the review groups apply a greater level of expertise and more rigorous assessment than others. These issues were identified to be common in Medical Technologies and Technology Appraisals, and could also be relevant to other programmes, particularly those programmes where the method manual guidance on assessing evidence is less well formed.

3.2. DISCUSSION

The specific challenges faced in each programme do differ somewhat. The unified Guidelines manual, covering Public Health and Social Care has established clear methods guidance and on the whole senior staff and committee members appear to be generally happy with the assessment of observational evidence within these programmes. However, there may be some further guidance required on combining RCT evidence with observational evidence, and assessing the uncertainty surrounding treatment effects. In the Technology Appraisal programme, it is relatively rare for observational data to be used. Despite this, it may be useful to provide guidance on best practice applications of observational data for the situations that could potentially arise, where observational data would benefit decision making and were not covered by TSD17 (Faria, et al, 2015). Especially given the impetuous towards increasing use of RWD as part of managed access agreements, and the inevitable problems this will bring for providing unbiased estimates of relative effectiveness and cost-effectiveness.

Medical Technologies, Interventional Procedures and Diagnostics Assessment appear to be faced with the issues that trials are often small and non-randomised. There is a lack of guidance on how treatment effects should be established in these programmes, and this needs to be addressed. Providing more detailed guidance on the assessment of bias in published non-randomised studies and the estimation of treatment effects using observational data is likely to greatly benefit the Medical Technologies, Interventional Procedures and Diagnostics Assessment programmes. In addition to this, training may be required to ensure that the review groups are fully aware of observational study design and analysis methods.

Managed entry agreements are an upcoming challenge facing many of the programmes. Development of guidance will be required in response to this. Faria et al, (2015), Rubin (2008) and Hernan and Robins (2016) emphasise the importance of good study design to

mitigate bias in observational studies. For each evidence-based managed entry agreement, the study design, protocol and statistical analysis plan should be defined at the outset. It is essential that the study protocol defines the appropriate population and the requisite data to be collected on all relevant confounding factors. Data on a suitable comparator needs to be identified. Guidance on the application of appropriate statistical methods should also be developed.

4. CASE STUDIES

In this section, we present three case studies to provide contrasting examples of where NICE could use RWE in the different NICE programmes. These case studies were selected after discussions with senior NICE staff representing different NICE programmes and after examining a range of candidate examples. We aimed to select cases that illustrate situations where different types of RWD had been incorporated into an appraisal, or where it seemed there may be a case for additional incorporation of RWD, that represented a range of challenges typically faced across different NICE programmes. Case studies are:

i) TA304 - total hip replacement and resurfacing arthroplasty for end-stage arthritis of the hip prosthesis for total hip replacement,
ii) MTG18 - the MAGEC system for spinal lengthening in children with scoliosis
iii) TA299 - bosutinib for previously treated chronic myeloid leukaemia.

For each of the case studies, we outline the decision problem, describe the evidence submitted to NICE, highlight the challenges posed by the evidence and suggest ways that real world data was used or could have been used to strengthen the evidence.

4.1. PROSTHESIS FOR TOTAL HIP REPLACEMENT (THR)

4.1.1. NICE TA on alternative prostheses for total hip replacement: TA304

Total hip replacement (THR) and surface replacement are common surgical procedures. The global market for hip prostheses was estimated at $4.7 billion in 2010 (Hip Replacement Implants, 2011). There are a large number of prosthesis brands that are often grouped into cemented, cementless and "hybrid" types. Hybrid prostheses consist predominantly of cemented stems and cementless cups.

The remit of the NICE Technology Appraisal required two sets of comparisons; (1) a comparison between hip resurfacing arthroplasty (RS) and THR, (2) and also a comparison between different types of THR. The assessment group chose to define prostheses type according to their fixation mode and bearing surfaces (Marques, et al., 2016), as follows:

- CePoM (Cemented-cemented with a polyethylene-metal articulation)
- CeLPoM (Cementless-cementless with a polyethylene-metal articulation)
- CeLCoC (Cementless-cementless with a ceramic-ceramic articulation)
- HyPoM (Hybrid (cementless-cemented) with a polyethylene-metal articulation)
- CePoC (Cemented-cemented with a polyethylene-ceramic articulation)

For both elements of the decision problem, the NICE appraisal required accurate estimates of the relative effectiveness of prostheses on long-term revision rates. Ideally, an RCT (or synthesis of RCTs) would be available that randomised a sufficiently large number of individuals to the different prostheses, and followed them up for a long period of time. The available RCT evidence, however, had small sample sizes or short durations of follow up (Morshed *et al*, 2007). The assessment group therefore obtained individual patient level data from the National Joint Registry (NJR) for England and Wales. This registry holds information on patient characteristics, type of prosthesis and time to revision for hip replacements. For the economic model, the Assessment Group estimated time to revision by prosthesis type, and used parametric survival models to predict the long-term relative effect of prosthesis type on revision rate. The assessment group limited the analysis to the above types of THR prostheses which were the most common and comprised 62% of the NJR procedures (239,000 cases).

The appraisal demonstrates the feasibility of using evidence from a large, existing observational dataset to inform NICE's guidance in circumstances where trial data is of limited value. However, the case study also raises some important concerns for the use of RWE in NICE decision-making more widely.

The assessment group's submission to NICE recognised the potential for selection bias due to confounding when estimating the relative effects of the choice of the technology on the outcome of interest (in this case revision rate). For the comparison of RS and THR, a

propensity score matching approach was taken which recognised that the probability of receiving RS or THR differed according to age. For the comparison of different THR types, the base case analysis did not allow for any differences between the patients receiving the different devices. In sensitivity analyses, the analysis used stratification by age and gender to recognise that these characteristics differed by the device. In particular cementless prostheses tend to be offered to younger patients.

A major concern was that neither of these analyses allowed for all the observed differences between the patients in the comparison groups according, for example, to their baseline physical status, which was considered by clinical experts to be a potentially important confounder. Hence, the lack of adjustment for confounding factors in the comparison of different types of THR was seen by the committee as an important limitation:

"The Committee concluded that the Assessment Group's analysis of revision rates was consistent with published systematic reviews of trials, and controlled for some, but not all, potential confounders, notably activity level and comorbidities, and therefore uncertainty remained surrounding the relative revision rates between different types of prostheses." (Section 4.3.6 FAD)

A further potential concern was that the assessment group's report did not consider whether the choice of technology had an effect on patient-reported outcomes, in particular health-related quality of life (QoL).

The Assessment group's base case results comparing prosthesis types showed that one category of prosthesis ("cemented polyethylene cup with a ceramic head (cemented stem)") dominated the other types included in the life-time analysis, including cementless and hybrid. However, the NICE guidance did not recommend any particular hip prostheses type over another.

4.1.2. Alternative analysis of total hip replacement (THR)

An alternative approach to the evaluation of alternative types of device for THR was taken in a study by Pennington *et al* (2013). The Pennington *et al* (2013) study addressed a somewhat different decision problem to that defined in the NICE assessment group's report, and compared three types of prosthesis (cemented, cementless, and hybrid) for THR for patients with osteoarthritis. The study estimated the relative effects of these alternative types of

primary THR, on costs and revision rates, but also post-operative QoL. In order to obtain data necessary to meet these requirements, the study obtained linked data from three large individual-level datasets.

These were;

- The NJR in England and Wales, with data on patients who had a THR from 2003-2009.
- Hospital Episode Statistics (HES) which records all publicly funded operations in England, with data used from patients who had a THR from 1997-2009
- Patient Reported Outcome Measures (PROMs) database which collects outcome data (EQ5D-3L and Oxford Hip, pre- and 6 months post-op), with data used from 2008-2010.

Pennington *et al* (2013) chose to estimate the relative effect of alternative prostheses for THR on post-operative QoL by using PROMs data linked to the NJR and HES, unlike the NICE assessment report which assumed equal QoL, regardless of the type of prosthesis. Hence this study made a wider use of observational data than the NICE appraisal and may offer some insights into how the design of a non-randomised study can be improved to try to reduce the potential for confounding.

In the Pennington et al study, the target population for inclusion in the analysis was tightly defined to improve the comparability of patients receiving each type of prosthesis. The sample was defined according to PROMs/NJR/HES linked data and included patients who had an elective THR, between July 2008 and December 2010. The main additional exclusions were patients aged under 55 or over 84, patients without diagnosis of osteoarthritis, patients undergoing a hip resurfacing procedure and privately funded patients. A sample of 30,203 patients had data available under these criteria.

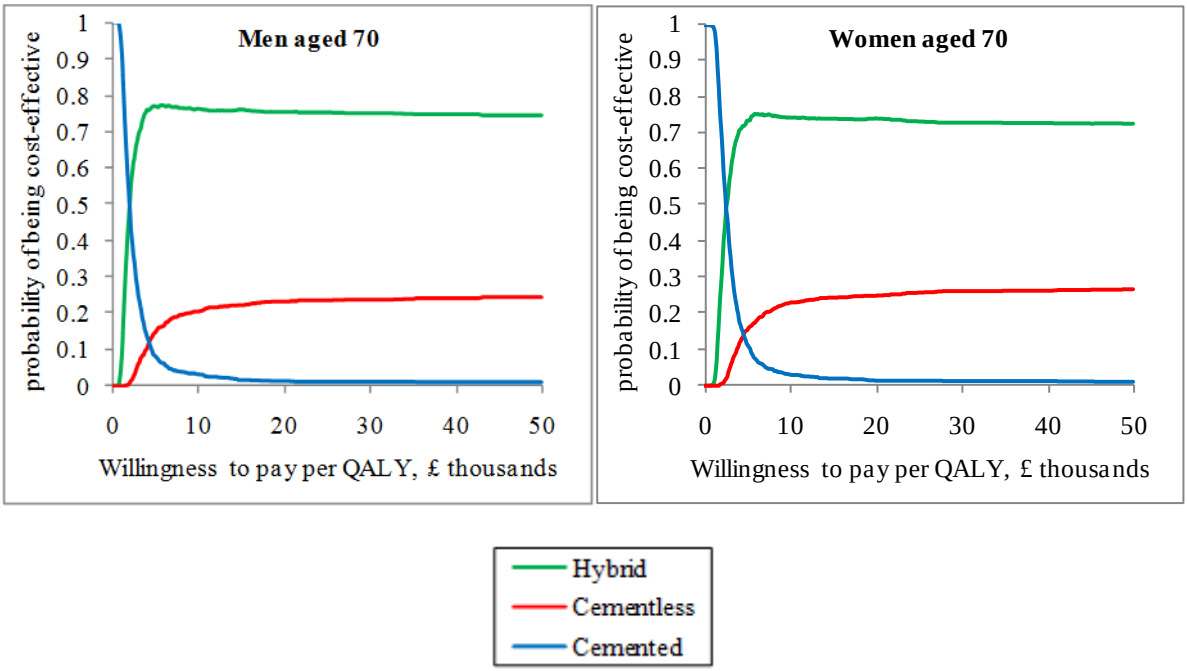
There was no obvious instrumental variable, so, as with the analysis undertaken in the NICE appraisal, Pennington *et al* (2013) used methods that assumed selection on observables. But a key difference was that here the study design attempted to ensure that selection on observables was a reasonable assumption. Specifically, for the estimations of the effect of prosthesis type on QoL, the access to the three linked observational datasets meant that case-mix information were available on patients baseline quality of life, activity levels and co-

morbidities, and the application of inclusion criteria meant that there was reasonable overlap on these prognostic characteristics. Any differences in the observed characteristics between the comparison groups were addressed using Genetic Matching, a multivariate matching method (Sekhon and Grieve, 2011). Patients were matched on age, sex, number of comorbidities, BMI, disability, physical status (ASA grade), pre-operative EQ-5D-3L and Oxford hip scores (OHS), surgeon experience (senior surgeon or not), and hospital type (treatment centre or not). The effect of prostheses type on quality of life was then estimated by applying an ordinary least squares (OLS) regression to the matched data. Regression adjusted matching was used because it has been shown to improve performance compared to relying on matching or regression alone (Kreif *et al*, 2012).

Similarly for the estimation of the effect of prosthesis type on revision rates up to five years, the study used NJR-HES linked data, but unlike the NICE assessment report adjusted for important confounders (age, sex, ASA grade, and BMI). for the estimated effect of prosthesis type on revision rates up to five years. The main advantage of the access to HES data was that it enabled the study to use observed data to predict revision rates beyond five years after hip replacement by fitting a survival model with a Weibull hazard function to patients with five to 12 years of follow-up data, adjusting for age and sex.

The results reported that QoL differed depending on which type of prosthesis is received, with the hybrid prosthesis leading to the highest mean EQ-5D-3L index score. The hybrid prosthesis also results in the highest Oxford hip score of the three alternatives. The results indicate that the 10 year revision rate was highest for those receiving a cementless, and lowest for those who had a cemented device inserted. The average cost of a cementless prosthesis was highest, the hybrid cost was slightly lower and cemented devices had the lowest cost of the three alternatives. The sensitivity analysis indicated that the results were robust after OLS was applied to the unmatched data, and according to alternative model specifications.

Figure 1: Cost-effectiveness acceptability curves for cemented, cementless and hybrid THR by subgroup plotted for alternative threshold willingness-to-pay for a QALY gained



The cost-effectiveness acceptability curves for each of the three prosthesis types and the main patient group (aged 70) are presented in Figure 1. The results show that the hybrid total hip replacement is relatively cost-effective and this result is largely driven by the finding that the hybrid prosthesis improves average QoL compared to cemented or cementless prostheses. The evidence submitted to NICE by the external review supported the use of cemented prostheses. Hence, the differences in the decision problem, but also the design of the Pennington *et al* (2013) study, appear to make a material difference to the results. The degree of adjustment for selection bias due to confounding was central to Pennington et al (2013). However, it should be noted that even in this study, the authors acknowledged that the cost-effectiveness estimates were still subject to the risk of unmeasured confounding.

4.1.3. Conclusions

The hip replacement case study illustrates a setting where different sources of RWE exist that are relevant to the appraisal, but not all of them are readily available. In the NICE submission the external review group used a single source of RWE, the NJR. By accessing the NJR data, the assessment group were able to provide estimates of long term revision rates, critical to the cost-effectiveness results. There are problems with estimating this parameter from RCTs with short follow-up periods. The situation where large scale registries of this type exist to complement trials may be relatively rare in the context of current Technology Appraisals, but

may be more frequently encountered in other NICE programmes and may become increasingly frequent even within the TA Programme. For example, the Systemic Anti-Cancer Therapies (SACT) dataset may provide richer routine data for certain chemotherapies. There is therefore a potential for additional information, collected in registries, that can complement trial data that are increasingly limited at the time of product licensing. Any increased use of RWD brings the challenge of minimising selection bias to provide meaningful and useful results.

Design of studies and the analyses undertaken can be crucial to the extent to which selection bias can be reduced (Hernan and Robins, 2016). The case study demonstrates that the study design, in this case the choice to include/exclude different sources of potentially relevant RWD, can determine key issues for the decision problem. This includes the range of comparators, as well as the range of potential confounders that are included/excluded from the subsequent analysis. In the Pennington et al (2013) study, great emphasis was placed on the inclusion/exclusion criteria for patients used in the analysis in order to ensure that there was sufficient balance between arms and to reduce issues raised by selection into treatment (for example according to age). This approach has some similarities to the restricted inclusion/exclusion criteria typically seen in clinical trials. As with an RCT, where an observational design applies inclusion and exclusion criteria, it is important to make this explicit and to assess how representative the sample of patients included is for the broader patient group for whom guidance would ideally be applied. A further issue raised by applying tight inclusion and exclusion criteria is that, this may limit the range of comparators that can be included in the CEA. One of the advantages of the Assessment Group analysis with respect to the TA decision problem was that it compared five different THR categories, compared to three in Pennington et al (2013). The greater aggregation of groups in the Pennington paper may have concealed important differences between THR sub-types (Jameson et al, 2015), although the authors did partly test this in sensitivity analyses, where for example they excluded patients with metal-on-metal prostheses and found similar results to the base case analysis.

The Pennington *et al* (2013) study illustrates the additional opportunities that may arise from accessing observational data linked across multiple sources, in this case data from a large registry (NJR) with those from administrative sources (HES and PROMs). A key advantage of accessing data from these multiple sources was that it enabled the study to adjust for a

broad range of prognostic differences between the comparison groups with multivariate matching and regression approaches. In contrast, the reliance on a single source of observational data hindered the extent to which the Assessment Group could allow for observed differences between the comparison groups. That said, even within the single observational data source, further adjustment could have been made for important potential confounders (for example, baseline physical activity). However, in any analysis that involves linking datasets, it is important to consider the additional biases that linking data can introduce into analysis. If the linkage rate is less than 100%, and missing data is not at random, it is necessary to consider how the associated reduction in sample size or imputation of missing values is likely to affect the results.

The Pennington *et al* (2013) study also used linked data to estimate other relevant parameters relating to cost and quality of life. It is interesting to note that notwithstanding the differences in the decision problems, that the NICE assessment group and the Pennington et al study chose to address, there are some differences in the main findings. The Pennington *et al* (2013) analysis reported that hybrid hip prostheses were the most cost-effective, whereas the base case analysis from the NICE Assessment Group indicated that cemented prostheses were likely to be cost-effective. These differences may be due to several reasons including the data used (particularly the inclusion of the PROMS data), the range of confounders included in the adjustment, and the ensuing analytical approach.

While there may be some advantages in the use of linked observational data, it may not be feasible within the current timescales for most NICE assessments. This would be dependent on the specific guidance programme. Furthermore, it is important to consider whether the exclusion of patients whose data cannot be linked will introduce bias into the analysis, and whether using for example multiple imputation approaches to handle these missing data would mitigate these biases.

Across the NICE programmes, it would be helpful to bring the decision making committee's attention to the sources of RWE that are available, the possibilities of access to individual patient data, and the opportunities for data linkage. At an early stage, it would be helpful to make the committee aware of the additional questions such an analysis might address, the extra time required, the extent to which analysts would anticipate that selection bias could be

mitigated if such work was undertaken, and the feasibility of undertaking such research within the planned timelines and budgets.

4.2. MAGEC

4.2.1. MAGEC submission

Early-Onset Scoliosis (EOS) refers to a spinal deformity which is present in children below the age of 10. In minor cases, the condition can be treated using bracing or casting. However, in more severe cases surgery is required. Around 120 children per year in England need this type of growth rod surgery. The surgery involves inserting growth rods into the back of the patient to control the progression of the spinal deformity as the child grows. Conventional growth rods need to be adjusted approximately every 6 months with an invasive procedure. There are of course complication risks and substantial impacts on patient and family wellbeing as a result of repeated surgery. In contrast, MAGEC rods can be lengthened using an external remote controller, and therefore claim to avoid repeated surgery. They therefore had a claimed potential to be substantially beneficial for patients but also to be cost saving to the NHS by reducing the number of surgeries children have to undergo (MTEP requires that positive guidance only be issued for technologies deemed to be cost neutral at least). The MAGEC system was appraised by the Medical Technologies Evaluation Programme (MTEP) in 2014.

There were a large number of outcomes of interest outlined in the scope, on which to compare MAGEC with conventional growth rods. Data of any type was particularly sparse and the sponsor (responsible for submitting the data submission to NICE in the MTEP programme) was only able to present evidence on:

- Total number of surgical procedures
- Total number of outpatient attendances and procedures
- Rate of distraction procedure success
- Infection rates and other surgical complication rates
- Device failure
- Change in Cobb angle (a measure of spine curvature) and spine height

The sponsor also presented evidence on pulmonary function and thoracic kyphosis, although these were not included in the meta-analysis. The available evidence consisted solely of non-

randomised (case-series and matched case series) studies, as there had not been any RCTs involving MAGEC.

The sponsor performed a systematic review of the literature for MAGEC. They identified three relevant published MAGEC case-series studies (Akbarnia *et al*. 2013a, Cheung *et al*. 2012, Dannawi *et al*. 2013), two relevant unpublished MAGEC case-series studies (Ellipse 2013, Yoon *et al*. 2013) and one unpublished matched case-series (Akbarnia *et al*. 2013b). According to the manufacturer, some of the patients appeared in more than one of the studies, therefore Yoon *et al*. 2013 and Akbarnia *et al*. 2013b were excluded to avoid double counting. A meta-analysis was performed on the remaining four studies to provide fixed-effects and random effects estimates for each available outcome. Outcomes were recorded at pre-operation baseline, post operation and follow-up. Mean follow-up times varied in each of the studies.

There was only one piece of evidence that compared MAGEC to conventional growth rods. The matched-case series study (Akbarnia *et al*. 2013b) compared a MAGEC case series with a matched comparison arm constructed from registry data. 17 patients were given MAGEC rods, 5 patients had missing data and were excluded from the sample and the remaining 12 patients were each matched to a patient from an Early-Onset Scoliosis (EOS) registry based on baseline characteristics, including gender, number of rods (single vs. dual), preoperative age, preoperative major curve (±20 degrees) and cause of scoliosis (congenital, idiopathic, neuromuscular, and syndromic). Details of the matching technique or algorithm were not included in the paper.

Table 4: Evidence included in the sponsor's submission and EAC report

Study	Device	Number of patients	Mean age of patients (years)	Mean follow-up time (years)	Mean change in Cobb angle	Mean number of surgeries per patient
Akbarnia *et al*. (2013a)	MAGEC	14	8.8	1.6	29*	1.1*
Cheung *et al*. (2012)	MAGEC	2	8.9	2	38.5*	1.0*
Dannawi *et al*. (2013)	MAGEC	34	8	1.3	28*	1.1*
Ellipse (2013)	MAGEC	30	7	1.8	22*	1.3*
Yoon *et al*. (2013)	MAGEC	6	7.5	-	-	-
Akbarnia *et al*. (2013b)	MAGEC arm	12	6.8	2.5	-	-
EACs meta-analysis or mean value	MAGEC	-	-	-	27.16 (FE) 27.17 (RE)	1.2

Akbarnia *et al*. (2013b)	CGR matched	12	6.6	4.1	19	6.5[α]
Bess *et al*. (2010)	CGR	140	6	5	28	6.1[α]
Andras *et al*. (2013)	CGR	37	-	4.1	35	7.0[α]
Caniklioglu *et al*. (2012)	CGR	25	7.3	6.6	31.6	5.4[α]
Farooq *et al*. (2010)	CGR	88	7	3.5	29.3[α]	-
Kabirian *et al*. (2012a)	CGR	402	5.9	6.3	-	-
Marquez *et al*. (2013)	CGR	24	-	3.7	28	-
McElroy *et al*. (2012)	CGR	27	7.6	4.8	36[α]	5.1[α]
Miladi *et al*. (2013)	CGR	23	9.3	3.5	39	2.8[α]
Pfandlsteine *et al*. (2012)	CGR	48	9.8	2.8	61	3.8[¥]
Sankar *et al*. (2011)	CGR	38	5.7	3.3	39[α]	-
Schroerlucke *et al*. (2012)	CGR	90	6	6	32.4	-
Uno *et al*. (2011)	CGR	39	8.5	4.2	45	-
Wang *et al*. (2012)	CGR	30	7.3	-	37.0[¥]	5.2[¥]
Watanabe *et al*. (2013)	CGR	88	6.5	3.9	32.3[α]	-
Zhao *et al*. (2012)	CGR	25	-	2.7	-	-
EACs meta-analysis (FE) or mean value <38 month follow-up	CGR	-	-	-	37.0	4.3
EACs meta-analysis (FE) or mean value >38wk month follow-up	CGR	-	-	-	32.14 (FE) 32.90 (RE)	5.8

*indicates study values used in EACs meta-analysis or mean calculations for MAGEC, [¥]indicates study values used in EACs meta-analysis or mean calculations for the <38 month follow-up CGR, [α]indicates study values used in EACs meta-analysis or mean calculations for the >38 month follow-up CGR. (FE) represents fixed effects and (RE) random effects.

Given that there were only 12 patients in the Akbarnia *et al*. 2013b conventional rod sample, the sponsor also identified another study (Bess *et al*. 2010) to provide evidence on conventional growth rods. Bess *et al*. 2010 was a published conventional rod case-series which included 140 patients. Mean values for each available outcome from the Bess *et al*. 2010 study and the conventional rod arm of the Akbarnia *et al*. 2013b were compared against the fixed-effects and random effects estimates from the MAGEC meta-analysis. See Table 4 for more details on the studies, including the number of patients included in each study, mean age of patient and mean follow-up time.

4.2.2. EAC report on MAGEC

A main concern for the external assessment centre (EAC) when interpreting the data on MAGEC compared to conventional growth rods, was selection bias due to confounding from the following factors; the underlying cause of the disease, age of onset, age of treatment, baseline Cobb angle, baseline spine and thoracic height, geographic location, type of

treatment received (single versus dual rods, differences in fixation techniques, additional bracing, surgeon performance, surgical and after-care protocols) and time to follow up. Furthermore, the EAC also suggested that the retrospective design of the Akbarnia et al. (2013b) study could lead to further bias because the inclusion criteria could be adjusted post-hoc to exclude patients with poor outcomes.

The EAC suggested that none of the MAGEC studies identified by the manufacturer took confounding factors into account either in the design or analysis of the study. However, Akbarnia *et al*. (2013b), attempted to take some of the confounding factors into account, but the quality of the matching cannot be assessed due to the lack of detail provided in the published paper. Summary statistics or balancing tests of the baseline characteristics were not presented in the paper.

The EAC suggested that one of the main weaknesses of the sponsor's submission was the lack of evidence on the comparator (Jenks, *et al*, 2014). Data on a comparator intervention was not collected as part of the MAGEC trials. Although Akbarnia et al. (2013b) presented comparator evidence using matched registry data, the patient numbers were very small (n=12 in each arm). The EAC undertook an additional review of the literature to identify studies detailing the performance of conventional growth rods. They identified a total of 16 relevant studies and a meta-analysis was performed by the EAC. As there were differences in follow-up times between the studies, the conventional rod studies were divided into two sub-groups depending on the length of follow-up time (less than 38 months and more than 38 months). Meta-analyses were performed on each sub-group. In addition to this, the meta-analysis on the four MAGEC studies was repeated, because the sponsor had not provided sufficient detail in their submission to assess the validity of their meta-analysis. Hence, analyses were performed by the EAC on the following sub-groups:

- Conventional rods – less than 38 months mean follow-up
- Conventional rods – more than 38 months mean follow-up
- MAGEC rods

4.2.3. Conclusions

The MAGEC case study illustrates a situation where there is a complete absence of RCT data and extremely limited observational data on either the intervention or comparators. It is a setting typical of many non-drug health technologies, which combines reliance on non-randomised data with several other characteristics (such as very small sample sizes) that make it difficult to provide unbiased estimates of treatment effects.

The problem of very small sample sizes affects the entirety of the evidence: studies on MAGEC in total (4 studies), studies on conventional growth rods (15 studies) and the individual studies that were used to make comparisons of effectiveness between MAGEC and conventional growth rods (1 study). In the latter case, matching was undertaken on just 12 patients.

MAGEC is also typical in terms of elements of process that are relevant here. The sponsor is the primary source of evidence and the analysis of that evidence that is placed in front of the decision making committee. There is an independent academic group that is responsible for providing a critique of that evidence and to a limited extent identifying and supplementing the analyses undertaken by the sponsor. Opinion from clinical experts and a patient representative were also considered in the decision-making process.

Despite these problems, the committee is required to make a comparison between conventional growth rods and the MAGEC system in terms of several outcomes and to establish the likely cost implications of adopting MAGEC for use in the NHS. The assessment of the cost case requires an assessment not only of whether MAGEC is superior or not to conventional growth rods in terms of reoperation rates, failure rates, duration of surgery and length of surgery *inter alia*, but an estimation of the magnitude of benefit in these outcomes.

The case study illustrates the problems with the construction of a dataset that appropriately represents the comparator (conventional growth rods) from historical data. The EAC rightly identified issues particularly in relation to the heterogeneity of conventional growth rods which may not represent current UK practice. It is essential that the comparator group is correctly identified, defined and data sourced on that group for valid comparisons to be made.

The lack of detail about the matching methods undertaken and the quality of the matches in the only attempt to make a comparison was a major concern. Whilst the very low numbers of patients identified for the MAGEC system will limit the number and range of potential confounders that can be balanced across the comparison groups, good practice would imply that the methods are made transparent and that the study tests the sensitivity of results to alternative methods (Kreif et al, 2013).

In this appraisal, this matched data played a critical role in the decision making process, yet the methods and quality of the matches were unknown. It is difficult to make a very detailed assessment of the relevance and robustness of the analysis undertaken, given that no summary statistics on the balance of the matched samples were presented in the Akbarnia *et al* (2013b) paper, and without access to the patient level data. At the time when MAGEC was selected for appraisal, it would have been clear that Akbarnia *et al* (2013b) was likely to take such an important role. The sponsor (Ellipse technologies) funded the key study (Akbarnia *et al*, 2013b) and it would have been reasonable to expect access to the patient level data, particularly since the authors had already published their analyses in a peer reviewed journal. At a minimum the access to individual patient data (IPD) in relation to the 12 patients that had received MAGEC, and the entirety (n=140) of the multicentre patient registry where controls were selected from would have been informative. But in addition, there would have been the possibility of constructing analyses that made use of the entirety of the evidence in relation to MAGEC rather than restricting comparisons to the 12 patients reported in Akbarnia *et al (*2013b). The sponsor funded three of the four studies used to provide evidence in relation to MAGEC.

Of course, there would still be major concerns about the comparability of data from such diverse sources, irrespective of the ability to conduct adjustments for selection on observables by gaining access to IPD and larger samples of patients. This type of analysis might give greater confidence in the plausibility of the cost-saving case and a means by which to judge threshold analyses on the magnitude of benefit required in order for MAGEC to remain cost saving. The value of on-going data collection on MAGEC in the NHS with a view to revision of the NICE recommendations, and how that data collection should be designed and analysed, could also be informed by analyses of IPD at the decision making stage.

4.3. BOSUTINIB FOR TREATMENT OF PREVIOUSLY TREATED PHILADELPHIA CHROMOSOME POSITIVE CHRONIC MYELOID LEUKAEMIA

4.3.1. Technology and decision problem

Philadelphia chromosome positive chronic myeloid leukaemia (CML) is a condition that is slowly progressive. Patients progress from the chronic phase to the accelerated phase and progress further to blast crisis phase (also called transformation). The incidence of CML in the UK lies between approximately 560 and 800 patients annually.

Bosutinib is indicated for the treatment of adult patients with chronic phase (CP), accelerated phase (AP) and blast phase (BP) Philadelphia chromosome positive CML, previously treated with one or more tyrosine kinase inhibitor(s) and for whom imatinib, nilotinib and dasatinib are not considered appropriate treatment options and for whom the only other option is hydroxycarbamide, that is considered to be best supportive care. Patients are treated with bosutinib until disease progression or until blast crisis (transformation) phase.

Bosutinib obtained a conditional marketing authorisation in January 2013. The marketing authorisation was and remains (at the time of this report) conditional on the company conducting a single arm open-label, multi-centre efficacy and safety study of bosutinib in patients previously treated with one or more tyrosine kinase inhibitors and for whom imatinib, nilotinib and dasatinib are not considered appropriate treatment options. The licence is to be reviewed in 2018 when first results from this research study are expected to become available. Initially, the manufacturer of bosutinib applied for marketing authorisation for first-line treatment of CML but a phase III study comparing first-line imatinib with first-line bosutinib had not shown superiority of bosutinib and the company therefore amended its application to the European Medicines Agency to the indication that is currently in use i.e. after failure of one or more tyrosine kinase inhibitors.

Bosutinib was appraised by NICE in September 2013 (TA299) and compared with hydroxycarbamide, considered to be equivalent to best supportive care.

The key issues in the appraisal were related to the evidence on bosutinib and its comparator. The evidence on bosutinib efficacy came from a subgroup of patients from Study 200 who received bosutinib 3rd line (N=118). Those that received bosutinib 3rd line but also for whom imatinib, nilotimib and dasatinib were unsuitable comprised a very small subgroup. Study

200 was a phase II single arm open label, efficacy and safety study of bosutinib once daily in patients with chronic, accelerated or blast phase CML. Overall survival in the analysis was 84% at 2 years.

Evidence on the comparator (hydroxycarbamide) was obtained from the subgroup who received hydroxycarbamide after imatinib failure in an observational cohort (N=12) (Kantarjian, 2007). The manufacturer estimated the mean overall survival for these 12 patients at 3.5 years. This is similar to the overall survival with hydroxycarbamide in a previous NICE appraisal on dasatinib, high-dose imatinib and nilotinib for the treatment of imatinib-resistant CML (TA241).

The Evidence Review Group (ERG) analysed the same data based on Kantarjian (2007) and estimated the mean overall survival as 7 years. This difference resulted from the company having cited TA251 ("Dasatinib, nilotinib and standard-dose imatinib for the first-line treatment of chronic myeloid leukaemia") that estimated the mean time on hydroxycarbamide treatment and assumed this to be equivalent to overall survival ignoring possible discontinuation in accelerated and blast phase. The ERG cited a study by Hoyle *et al*. (2011) that fitted an exponential model to the 2 and 3 years survival of 77% and 70%, respectively. The Committee, after consulting with clinicians, stated that the mean overall survival estimate of 3.5 years for patients treated with hydroxycarbamide was the most plausible.

Because of the lack of long-term evidence and the lack of comparative data, the modelling of overall survival was considered an important issue. The Committee considered the "cumulative" approach to estimating overall survival for bosutinib, that was suggested by the ERG, the most appropriate. Within this approach, the duration of treatment with bosutinib was added to the overall survival estimate for hydroxycarbamide to yield the overall survival estimate for bosutinib. The bosutinib treatment duration was estimated using a log-normal curve fitted to discontinuation data from Study 200.

The Committee did not recommend bosutinib based on ICERs that were not within the acceptable range (estimated to be between £43,000 to £49,000 per QALY for the chronic phase). The key uncertainties were related to the lack of evidence on the efficacy of bosutinib and hydroxycarbamide in terms of overall survival and other outcomes.

4.3.2. Potential use of RWD in appraisal

When bosutinib was appraised in 2013, NICE did not recommend further research within an "Only in Research" or "Recommendation with Research" scheme. These schemes would have promoted further collection of data for consideration at a defined time of re-appraisal. Such a scheme could potentially have been used to provide more accurate estimates of the relative treatment effect of bosutinib compared with hydroxycarbamide. Since outcomes are uncertain with both treatments, a randomised controlled trial with follow up long enough to capture effects on overall survival would clearly have been the most appropriate method for estimating the survival benefit. However, the very small patient population may mean that such research may not be associated with a positive expected net gain when comparing the value of research with its cost. These considerations are important and require formal assessment when making research recommendations.

Since the appraisal in 2013, bosutinib was reimbursed for patients with previously treated Philadelphia chromosome positive CML through the Cancer Drugs Fund (CDF). In light of the changes being made to the CDF, bosutinib is currently being re-appraised by NICE with an anticipated publication date of October 2016 (see NICE website https://www.nice.org.uk/guidance/indevelopment/gid-ta10040). The use of bosutinib via the CDF has provided an opportunity for data to be collected from NHS patients in the interim. Whether that data has been collected and forms part of the re-appraisal remains to be seen. However, two issues remain: a) a relatively long follow up is needed to obtain better data on overall survival; b) observational data on bosutinib alone does not resolve the likely bias in any estimates of survival gain compared to the comparator hydroxycarbamide because of the scarcity of data on this comparator. In the absence of the opportunity for retrospective analysis of registry data, the funding of bosutinib in the entire eligible NHS population closed down the opportunity to resolve this aspect of uncertainty. Both issues a) and b) could have been addressed by a recommendation that was conditional on additional data collection being undertaken. An alternative could be to wait for the reporting of the single-arm study that is currently underway and expected to report in 2021, though again as this is a single arm study this will also not resolve any uncertainty about the survival gains compared to hydroxycarbamide.

4.3.3. Conclusions

In the period between bosutinib being rejected by NICE for routine NHS use, its subsequent adoption by the CDF and the present time there has been the opportunity to collect data from NHS patients. Current proposals for changes to the CDF advocate adoption on to the CDF in those circumstances where additional data collection will resolve uncertainties within a 2 year period, with an expectation that the drug is likely to be cost effective once that uncertainty is resolved. This case study suggests that such data could have been, and in part was, collected. Yet it also serves to illustrate the limitations of such additional real world data, the importance of adoption decisions, and the many challenges faced.

The manufacturer submission indicated that the 2 year survival for patients treated with other treatments (amongst them patients that had received hydroxycarbamide) was 77%. Therefore short follow up of patients, irrespective of the study design, is likely to be a substantial limitation to any robust estimate of overall survival gain from bosutinib. This may be magnified in a scenario where there are additional concerns about the biases associated with non-randomised data.

The other major limitation in this case study is the extremely limited basis on which any comparison with NHS practice could have been made. The committee were of the opinion that the appropriate comparator was hydroxycarbamide, yet data was so scant on this as to comprise 12 patients within a sample of 61 that had received "other treatments". The funding of bosutinib through the CDF made it impossible to improve upon this very poor evidence base with additional real world evidence collection. It is worth noting that NICE made no recommendation to collect additional data on this comparator so it is not clear if this would have occurred even without the actions of the CDF.

The case illustrates a risk associated with early access to medicines where there is a lack of data on the outcomes of similar patients who receive the relevant comparator(s). The availability of a new therapy to UK patients may make it impossible to reduce the area of greatest uncertainty in its assessment. NICE's recommendations regarding additional research must consider the data required on comparators as well as new therapies, particularly when there is no randomised comparison.

5. SUMMARY OF OVERALL RECOMMENDATIONS

Across all NICE programmes, we have identified a significant reliance on non-randomised data in formulating guidance. This occurs with varying frequencies across the programmes but in all cases there are different, additional challenges presented compared to situations where the guidance is directly informed by RCT evidence. There are potential changes to NICE methods and processes that may help the Institute to improve the way in which such data is incorporated into its assessments, and to align best practice across its activities. Such changes are likely to become increasingly important as initiatives to speed up the process of patient access to new technologies grow in influence. Without this continued progress, NICE risks issuing guidance based on biased assessments of the true effectiveness of interventions.

This review highlighted the importance of defining the relevant counterfactual to the intervention of interest. This is a key issue across the programmes of interest. For example, the major limitation in case studies 2 and 3 was that there was no clear definition of the relevant comparator/counterfactual, and hence no attempt was made to collate appropriate data on the case-mix and outcomes of a relevant comparator. Unless, the relevant comparator is defined at the outset, then this makes any subsequent attempt to allow for potentially confounding factors at the analysis stage problematic, no matter how sophisticated the statistical method, Rubin (2008), Hernan and Robins (2016).

The availability of individual level data on baseline (prognostic) characteristics and outcomes, offers greater potential to undertake subsequent analyses that attempt to minimise the inevitable treatment selection bias in non-randomised studies. By contrast, where there are only summary level data available for the comparators of interest, for example for pre-treatment prognostic characteristics, this limits the scope of studies to fully adjust for selection bias. It is typically the case that methods for adjusting for selection bias identified as a result of the lack of randomisation can best be interrogated, and alternatives examined, with full access to data at the patient level. This is true even in those situations where sample sizes may be very small (as, for example, in the MAGEC case study), where there may be limited scope for applying different quantitative approaches to address potential biases. Even in these situations there is the scope to examine the matching procedure used and the sensitivity of results to alternative matching approaches. Hence, a strength of the Pennington *et al* (2013) case study was that an extensive sensitivity analysis was undertaken which involved testing

the extent to which the overall conclusions were robust to the choice of analytical approach (matching and regression versus either matching or regression alone), and also to the particular specification of the parametric model.

Dealing with non-randomised data to estimate treatment effects, and being qualified to critique analyses undertaken by others, requires specialist skills that do not necessarily reside in the groups responsible for evidence submission or critique in NICE's processes. Methods typically required are not routinely well understood even by those skilled in data analysis. There will be variation between existing groups in their abilities to source appropriate expertise. There could be benefits from additional resources from NICE and other sources to facilitate more widespread knowledge of appropriate methods: changes to NICE's methods guides will almost certainly need to be accompanied by more technical guidance (see the DSU Technical Support Document 17 on observational data methods (Faria *et al*, 2015) that covers some of this), associated educational activities and checklists outlining commonly agreed good practice. Currently, NICE methods guides provide little guidance in relation to how to critique analyses of estimates of treatment effects based on non-randomised data. Some of the guides have detailed instructions which focus on different study designs and their associated sources of bias, but they say little about analyses that attempt to deal with those biases. As a consequence, typically there is a thorough analysis provided to advisory committees of the strengths and weaknesses of the various study designs where clinical evidence is drawn from. None of the methods guides that were reviewed provide direct advice to those designing or analysing the primary non-randomised study on how to minimise selection bias. As NICE moves to greater use of non-randomised evidence, this will pose further issues for the available methods, beyond those advocated in previous guidelines. In particular, while the Faria et al 2015 TSD deals with methods appropriate for handling confounding in treatments delivered at a particular point in time, further research is required on methods that handle confounding factors that change over time. This particular form of confounding has to be addressed in providing treatment recommendations for chronic diseases where sequential treatment decisions may depend upon patient's prognosis which changes over time.

Where access to patient level data can be obtained, there is also likely to be a requirement for academic groups to obtain research governance and ethics approval in order to receive that data (even though it would be expected to be anonymised). This, together with the challenges

outlined above, make it critical that those NICE assessments likely to rely substantially on non-randomised data are identified at a very early stage of the assessment. An explicit step for NICE at the stage of scoping work for an assessment should be to consider this issue. In most situations the lack of randomised data would be known *a priori*. In this situation NICE should establish that there is an expectation for access to patient level data to be sought. That expectation can be relayed to those that submit evidence both at the scoping stage of the assessment but more generally the Institute could work to establish other means of conveying this message, for example through its Scientific Advice programme.

In particular, for those assessments where manufacturers are the data holders or sponsors of studies, it may be feasible for appropriate interrogation of data to be undertaken within existing timelines. In other settings it will not always be feasible to obtain data in a timely manner, in which case judgments about the additional value such access would bring must be made.

There are already many scenarios where NICE may make decisions that are to some extent dependent on future research, and this is likely to expand. In its crudest form this may be simply by offering research recommendations with no formal link to future decision making, but increasingly there is a need for this link to be explicit. The case studies demonstrate the need to consider at a very early stage the design and analysis of non-randomised studies intended to make unbiased estimates. Without adequate consideration of how potential biases will be addressed, decision making committees will feel a greater degree of uncertainty about the decisions available to them and the associated risks to the NHS.

Ideally, such research recommendations would be accompanied by an explicit, quantitative assessment of the burden uncertainty constitutes and the extent to which that NHS burden is reduced through additional research (see DSU report on Managed Entry Agreements, Grimm *et al*, 2016). Even without that quantitative assessment, issues of design, for example where comparator data are collected from, length of follow up, outcomes, patient selection, heterogeneity, sample sizes, all need to be considered at the outset to avoid misallocation of research funds and inappropriate decisions being made. The methods by which these design decisions can be taken may differ depending on whether there is the opportunity to collect data on all interventions or only on the new technology. The hip replacement example illustrates decisions that were taken in order to make the treatment groups that were included

in the analysis comparable. That involves a trade-off with generalisability that can sometimes be the very reason motivating the use of RWD in the first place.

The value of NICE methods guidance on the design, analysis and use of non-randomised data would be of particular value to those setting up studies in this context. Furthermore, the example of bosutinib illustrates the importance of data on the comparator and, whilst NICE may face pressure to grant access to NHS patients even in the face of substantial risk, the implications of routine access in gathering data on the comparator need to be explicitly incorporated into the evaluation of any risk management scheme. Unless careful consideration is given to the counterfactual comparator, the use of RWD will lead to inaccurate estimates of treatment effects.

The hip replacement case study shows the potential value that large datasets can offer to decision making and how sophisticated methods that attempt to overcome potential bias from confounding can be used to obtain estimates of outcomes. An advantage of accessing linked data (e.g. NJR-HES), was that it allowed the analysis (e.g. of the effect of different prosthesis types on revision rate) to adjust for a wider range of measured confounders than if the analysis had relied on a single data source. So, while baseline information on age and patient activity were available from a clinical database (the NJR), additional baseline information, on the number of comordities was only available via linkage to an administrative dataset (HES). These richer datasets are becoming typical of many specific disease areas and general health services alike. The Clinical Practice Research Datalink (CPRD), HES, death certificates, cancer registries *inter alia*, and their ability to link records offer huge potential benefits for Health Technology Assessment (HTA). The ability to access such rich data sources, to link them together, and analyse them raises the question as to whether bodies like NICE ought to reassess the way in which the traditional hierarchy of evidence has translated into processes. Typically, systematic reviews of evidence on treatment effectiveness would be restricted to trial based studies, only venturing to lower levels of evidence where randomised evidence is limited.

The use of these rich, linked datasets may also raise challenges that require further research, or at least more guidance. Care must be taken to ensure that the linkage of data does not create sample bias by dropping observations that cannot be linked. The issue of time varying confounders and the ability to employ methods to adjust for such confounding become

feasible in these types of analysis. They also allow decisions to be made about the appropriate design and selection of the study sample in order to optimise the comparisons of patients receiving the treatment versus those receiving the control. Because smaller datasets may not offer the same scope to employ methods that potentially adjust for these biases, they are often ignored.

In many situations we see comparisons being drawn between patients that receive an intervention in one study and patients receiving comparators in other studies. This is clearly the case when using single-arm studies. In response to this, methods work must be undertaken and subsequent guidance produced to provide best practice advice for situations where there are disconnected evidence networks, including advice on how to analyse data when faced with aggregate data from published studies, and how to select historical cohorts to ensure compatibility with contemporary single-arm studies.

Box 1: Summary of main priorities for consideration

Priority issues for NICE methods and processes

1. Where a reliance on non-randomised data to inform decision making is identified, access to individual patient level data (IPD) for analysis and independent review is advantageous. NICE should signal to its stakeholders that there is an expectation of access to IPD where feasible. Changes to processes may be required to:
    a. Identify assessments where this is the case at the earliest opportunity
    b. Permit the timely incorporation of IPD.
2. NICE should produce, or help to commission and endorse, guidance, consistent across programmes, on methods for design, analysis, and interpretation of non-randomised evidence. Consideration should be given to the skills required and training needs of NICE staff and stakeholders in this area, particularly if a greater use of this type of evidence is promoted.
3. In those situations where options that link decision-making to future research are considered, NICE should seek to develop detailed instructions on all stages of that future research and how it would reduce current decision uncertainty. This should cover the design, planned analysis and incorporation of the results into cost-effectiveness models.
4. NICE should consider whether large nationally funded datasets, and the potential for data linkage, be routinely examined as part of its appraisal process even where randomised evidence may exist.

Priority research requirements

1. The production of a comprehensive manual for the design, analysis and interpretation of results from observational studies into decision making for use in managed entry agreements.
2. Research is needed into the extent to which observational designs can complement or substitute those of RCTs in resolving the biases and uncertainties typically encountered. This is particularly important in relation to the reformed Cancer Drugs Fund, and may have relevance to other areas of NICE work, including the assessment of medical diagnostics and devices.
3. A review of existing methods for the identification of and adjustment for time-varying confounders and the development of new methods, if required, for use in Health Technology Assessment.
4. Investigation into methods for estimating treatment effects when data is drawn from intervention patients in one setting versus comparator patients from a different setting (according to time or location). Methods need to distinguish between scenarios where there may be IPD or aggregate level data (or both) on one or more of the relevant comparators.

REFERENCES

Akbarnia BA, Cheung K, Noordeen H, Elsebaie H, Yazici M, Dannawi Z, et al. Next generation of growth-sparing techniques: preliminary clinical results of a magnetically controlled growing rod in 14 patients with early-onset scoliosis. Spine. 2013;38(8):665–70.

Akbarnia BA, Cheung K, Noordeen H et al. Traditional growing rods versus magnetically controlled growing rods in early onset scoliosis: a case-matched two year study. 2013. https://www.growingspine.org/news-and-events/growing-spine-study-grouppapers-presented-at-the-imast-and-srs-annual-meeting. Accessed 10 July 2014.

Andras L, Joiner E, McCarthy RE, et al. Early onset scoliosis treated with growing rods has a greater increase in T1-S1 length, better Cobb correction and more than twice the number of surgeries compared to Shilla. In: Scoliosis Research Society 48th Annual Meeting and Course; 18–21 September 2013; Lyon.

Bess S, Akbarnia BA, Thompson GH, Sponseller PD, Shah SA, El Sebaie H, et al. Complications of growing-rod treatment for early-onset scoliosis: analysis of one hundred and forty patients. J Bone Joint Surg Am. 2010;92(15):2533–43.

Caniklioglu M, Gokce A, Ozturkmen Y, Gokay NS, Atici Y, Uzumcugil O. Clinical and radiological outcome of the growing rod technique in the management of scoliosis in young children. Acta Orthop Traumatol Turc. 2012;46(5):379–84.

Cheung KM, Cheung JP, Samartzis D, et al. Magnetically controlled growing rods for severe spinal curvature in young children: a prospective case series. Lancet. 2012;379(9830):1967–74.

Craig P, Dieppe P, McIntyre S et al. on behalf of the MRC (2008) Developing and evaluating complex interventions: new guidance. London: Medical Research Council

Craig P, Cooper C, Gunnell D et al. on behalf of the MRC (2011) Using natural experiments to evaluate population health interventions: guidance for producers and users of evidence. London: Medical Research Council

Dannawi Z, Altaf F, Harshavardhana NS, El Sebaie H, Noordeen H. Early results of a remotely-operated magnetic growth rod in early-onset scoliosis. Bone Joint J. 2013;95-B(1):75–80.

Ellipse Technologies Inc. A retrospective multicenter review of early onset spinal deformity patients that underwent either a primary or revision spinal bracing procedure with the ellipse technologies MAGEC spinal bracing and distraction system. 2013. http://clinicaltrials.gov/show/NCT01716936. Accessed 10 July 2014.

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

Farooq N, Garrido E, Altaf F, Dartnell J, Shah SA, Tucker SK, et al. Minimizing complications with single submuscular growing rods: a review of technique and results on 88 patients with minimum two-year follow-up. Spine. 2010;35(25): 2252–8.

Final appraisal determination (FAD) TA304 – Total hip replacement and resurfacing arthroplasty for end stage arthritis of the hip (review of technology appraisal guidance 2 and 44) https://www.nice.org.uk/guidance/ta304 (Accessed: May 2016)

Grimm, S, Strong, M, Brennan, A, Wailoo, A.J. Framework for analysing risk in Health Technology Assessments and its application to Managed Entry Agreements. NICE DSU report. 2016 Available from http://www.nicedsu.org.uk

Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. American Journal of Epidemiology. 2016 Apr 15;183(8):758-64.

Higgins, J. P. T., and S. Green. "Cochrane handbook for systematic reviews of interventions version 5.1. 0 [updated Mar 2011]". The Cochrane collaboration.

Hip Replacement Implants - Global Pipeline Analysis, Competitive Landscape and Market Forecasts to 2017. Rockville, MD: MarketResearch.com, 2011.

Hoyle M, Pavey T, Ciani O, et al. Dasatinib, Nilotinib, and standard dose Imatinib for the first-line treatment of chronic myeloid leukaemia: systematic reviews and economic analyses. National Institute for Health and Clinical Excellence (NICE): Peninsula Technology Assessment Group (PenTAG), University of Exeter 2011.

Jameson, S.S., Mason, J., Baker, P.N., Gregg, P.J., Deehan, D.J. and Reed, M.R., 2015. Implant optimisation for primary hip replacement in patients over 60 years with osteoarthritis: a cohort study of clinical outcomes and implant costs using data from England and Wales. PloS one, 10(11), p.e0140309.

Jenks, Michelle, Joyce Craig, Joanne Higgins, Iain Willits, Teresa Barata, Hannah Wood, Christine Kimpton, and Andrew Sims. "The MAGEC system for spinal lengthening in children with scoliosis: a NICE Medical Technology Guidance." Applied health economics and health policy 12, no. 6 (2014): 587-599.

Kabirian N, Akbarnia BA, et al. Deep surgical site infection following growing rod surgery in early onset scoliosis: how does it change the course of treatment? In: Scoliosis Research Society 47th Annual Meeting and Course; 5–8 September 2012; Chicago.

Kantarjian H, O'Brien S, Talpaz M, et al. Outcome of patients with Philadelphia chromosome-positive chronic myelogenous leukemia post-imatinib mesylate failure. Cancer 2007;109:1556-60.

Kreif N, Grieve R, Radice R, Sadique Z, Ramsahai R, Sekhon JS. Methods for estimating subgroup effects in cost-effectiveness analyses that use observational data. Med Decis Making [Internet]. 2012 16/10/12. Available from: http://mdm.sagepub.com/content/early/2012/06/11/0272989X12448929.

Kreif N, Grieve R, Sadique MZ. Statistical Methods For Cost-Effectiveness Analyses That Use Observational Data: A Critical Appraisal Tool And Review Of Current Practice. Health economics. 2013 Apr 1;22(4):486-500.

Marques, E.M., Humphriss, R., Welton, N.J., Higgins, J.P., Hollingworth, W., Lopez-Lopez, J.A., Thom, H., Hunt, L.P., Blom, A.W. and Beswick, A.D., 2016. The choice between hip prosthetic bearing surfaces in total hip replacement: a protocol for a systematic review and network meta-analysis. Systematic reviews, 5(1), p.1.

Marquez S, Miguel J, Francisco Javier SPG, Alfredo GF, De La Sacristana Nicomedes FBG, Jose QR. Reconstructive surgery in patients with early onset scoliosis (EOS) treated with growing rods. Eur Spine J. 2013;22(1):213–4.

McElroy MJ, Sponseller PD, Dattilo JR, Thompson GH, Akbarnia BA, Shah SA, et al. Growing rods for the treatment of scoliosis in children with cerebral palsy: a critical assessment. Spine. 2012;37(24):E1504–10.

Miladi L, Journe A, Mousny M. H3S2 (3 hooks, 2 screws) construct: a simple growing rod technique for early onset scoliosis. Eur Spine J. 2013;22(Suppl 2):S96–105.

Morshed S, Bozic KJ, Ries MD, Malchau H, Colford Jr JM. Comparison of cemented and uncemented fixation in total hip replacement: A meta-analysis. Acta Orthopaedica. 2007;78(3):315-26.

National Institute for Health and Care Excellence. The MAGEC system for spinal lengthening in children with scoliosis. 2014. http://www.nice.org.uk/guidance/MTG18. Accessed 10 July 2014.

NICE. Developing NICE guidelines: the manual http://www.nice.org.uk/article/pmg20 (Published: 31 October 2014 updated: 21 January 2016 Accessed: 9 February 2016)

NICE. Diagnostics Assessment Programme manual (Issue date: December 2011) Interventional procedures programme manual http://www.nice.org.uk/article/pmg28 (Published: 01 February 2016 Accessed: 9 February 2016)

NICE. Methods for the development of NICE public health guidance (third edition) http://www.nice.org.uk/aboutnice/howwework/developingnicepublichealthguidance (Published: 26 September 2012 Accessed: 9 February 2016)

NICE. The social care guidance manual http://publications.nice.org.uk/pmg10 (Published: 30 April 2013 Accessed: 9 February 2016)

NICE. The guidelines manual. https://www.nice.org.uk/article/pmg6/chapter/1%20Introduction (Published: November 2012 Accessed: 24 May 2016)

NICE. Medical Technologies Evaluation Programme - Methods guide https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-medical-technologies-guidance (Issue date: April 2011 Accessed: 9 February 2016)

NICE. Guide to the methods of technology appraisal http://publications.nice.org.uk/pmg9 (Published: 2013 Accessed: 9 February 2016)

Pennington, Mark, Richard Grieve, Jasjeet S. Sekhon, Paul Gregg, Nick Black, and Jan H. van der Meulen. "Cemented, cementless, and hybrid prostheses for total hip replacement: cost effectiveness analysis." (2013): f1026.

Pfandlsteiner T, Seidel K, Wimmer C. Growth modulation to continue spinal growth in juvenile scoliosis: 8 year follow up. Eur Spine J. 2012;21(11):2326.

Pisani, J, Lee, M. Accelerated Access Review Proposition 2: Getting ahead of the curve - Recommendations for accelerated access pathways and a flexible pricing and reimbursement framework, https://www.gov.uk/government/publications/accelerated-access-pathways-for-medical-technologies [accessed 5th May 2016]

Rubin DB. For objective causal inference, design trumps analysis. The Annals of Applied Statistics. 2008 Sep 1:808-40.

Sankar WN, Skaggs DL, Yazici M, Johnston CE 2nd, Shah SA, Javidan P, et al. Lengthening of dual growing rods and the law of diminishing returns. Spine. 2011;36(10):806–9.

Schroerlucke SR, Akbarnia BA, Pawelek JB, Salari P, Mundis GM Jr, Yazici M, et al. How does thoracic kyphosis affect patient outcomes in growing rod surgery? Spine. 2012;37(15):1303–9.

Sekhon JS, Grieve R. A matching method for improving covariate balance in cost-effectiveness analyses. Health Econ. 2011;21(6):695-714.

Uno K, Suzuki T, Kawakami N, et al. The effect of early fusion at ten years or earlier for early onset scoliosis: comparison between 43 early fusion patients and 39 growing rod patients. In: Scoliosis Research Society 46th Annual Meeting and Course, 7–14 September 2011; Louisville.

Wang SZJ, Qiu G, et al. Dual growing rods technique for congenital scoliosis: more than 2 years outcomes. Preliminary results of a single center. Spine. 2012;37(26):E1639–44.

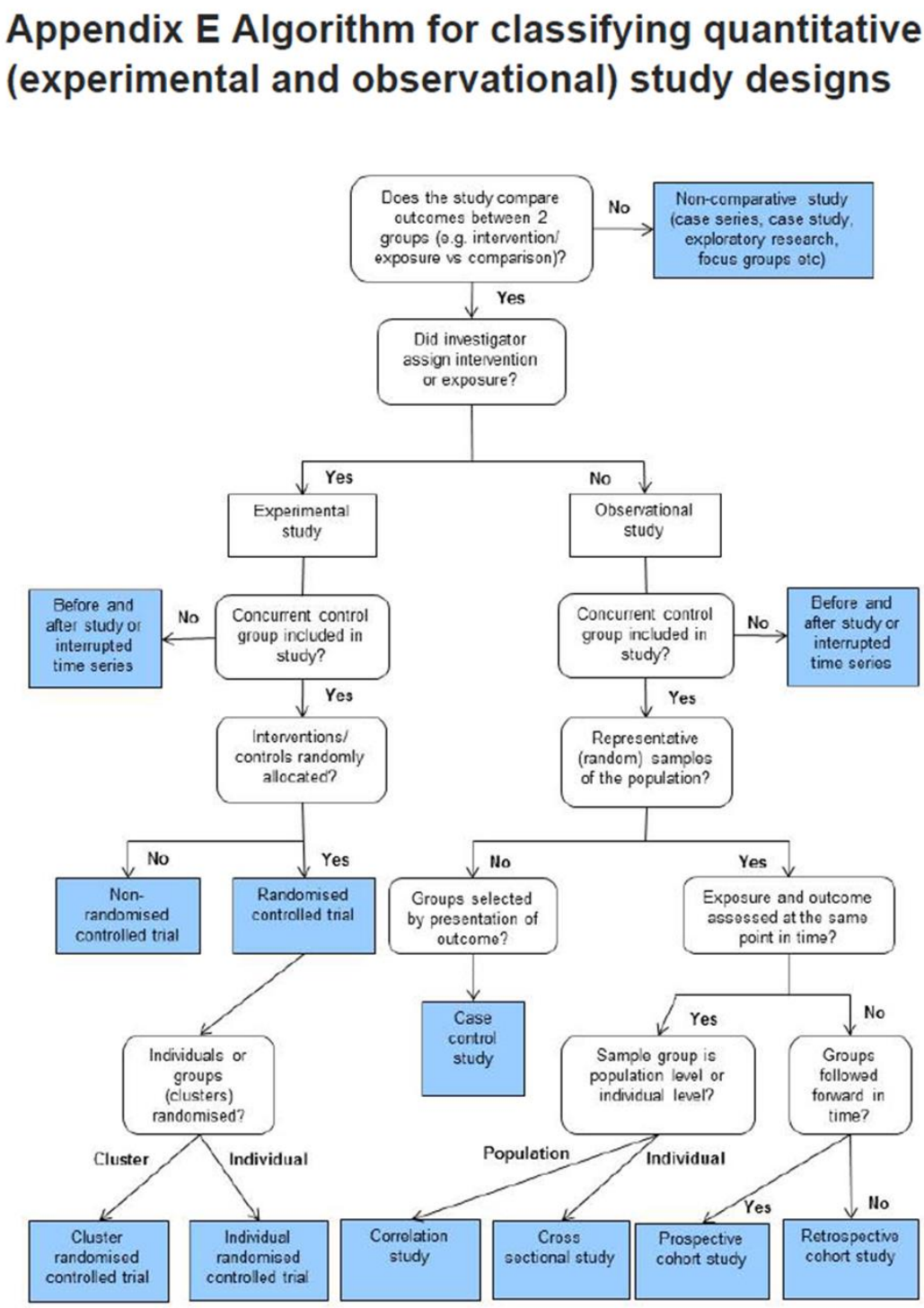
Watanabe K, Uno K, Suzuki T, Kawakami N, Tsuji T, Yanagida H, et al. Risk factors for complications associated with growing rod surgery for early-onset scoliosis. Spine. 2013;38(8):E464–8.

Victora C, Habicht J, Bryce J (2004) Evidence-based public health: moving beyond randomized trials. American Journal of Public Health 94(3): 400–5

Zhao Y, Qiu G-X, Wang Y-P, Zhang J-G, Shen J-X, Li S-G, et al. Comparison of initial efficacy between single and dual growing rods in treatment of early onset scoliosis. Chin Med J. 2012;125(16):2862–6.

Yoon WW, Sedra F, Suken S, et al. Improvement of pulmonary function in children with early onset scoliosis using magnetic growth rods. Spine. 2014;39(15):1196–202.

APPENDIX 1: PUBLIC HEALTH AND GUIDELINES - QUANTITATIVE STUDIES FLOW CHART



Source: NICE Public Health and Guidelines Manuals