



Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Matching-Adjusted Indirect Comparisons: Identifying Method Variations and Implementing Models in R

A thesis submitted to the University of Dublin, Trinity College in fulfilment of the requirements for the degree of Master in Science (Research) in Statistics

2020

Owen Cassidy

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish.

Owen Cassidy

Abstract

In the framework of evidence-based medicine, comparative effectiveness research is a fundamental activity to the development of pharmaceutical products and medical treatments. For a given medical condition, several competing treatments may exist, however, there may be no multi-arm study available comparing all the relevant treatments simultaneously. In the absence of such direct evidence, healthcare decision-makers can examine indirect comparisons of treatments generated using outcome data from separate clinical trials. In order to theoretically reduce the possibility of error in the estimates of treatment effect for each treatment obtained from the indirect comparison, adjustments can be made to account for differences between trials. Matching-Adjusted Indirect Comparisons (MAICs) have been devised as a means of accounting for the heterogeneity that can arise between different patient populations in separate trials. Using individual patient data (IPD) from one trial, baseline patient characteristics and patient outcomes are weighted to match to published aggregate summary data available for a competing treatment. This statistical technique is relatively new and the various implications for its use have not been fully explored. In addition, while the technique has been employed in numerous published applications, the finer details of its implementation are rarely fully reported. The work in this thesis aims to articulate a greater understanding of the MAIC method and standardise its implementation through means of a package for the R programming language, which will allow for analyses to be reproducible.

Summary

The aim of this thesis is to establish and articulate an understanding of population-adjusted treatment comparison methods and in particular, assess the potential impact of Matching-Adjusted Indirect Comparisons (MAIC) on Health Technology Assessment (HTA).

In order to lay the foundation to the methodology and motivation for MAIC, some key underlying concepts are explored. These include the framework of evidence synthesis, epidemiological measures relating to treatment effect and challenges in making inferences on treatment effectiveness based on clinical trial data. The concept of population adjustment is introduced and the logistic regression statistical model is presented and linked to MAIC. Following this, a review of the MAIC methodology literature was conducted and an outline of the MAIC method is described based on this review. Subsequent to this, a separate review of publications making use of the method is detailed. Both searches were conducted using the PubMed platform and Embase database. Together with the review of MAIC methodology, it becomes clear that a standardised implementation of the method is lacking. This has an impact on HTA, since without a standardised set of parameters and protocol to draw upon, various health technology submissions may implement the method through one of a multitude of ways. This can lead to variation in results and raises the risk of unequal assessment between different technology appraisals.

From the literature review, this thesis seeks to identify best practice with respect to implementing MAIC and establish a standardised implementation of the method using the R programming language. This is preceded by a search of R packages available on the Comprehensive R Archive Network (CRAN) at the time of review. Some packages of relevance to propensity score weighting and network meta-analysis are identified. This further motivates development of the R package for implementing MAIC.

Development of the R package is carried out in accordance with best practice measures published by Hadley Wickham. The package is based on example code provided in a Technical Support Document (TSD) produced for the National Institute for Health and Care Excellence (NICE). The final design allows for a user to match one or more binary or continuous covariates from a set of individual patient data (IPD) to a proportion (for binary covariates) or mean and standard deviation (for continuous covariates). With additional input, the package also allows the user to perform the indirect comparison for a binary outcome measured on the log odds scale. Using the population-adjusted treatment effect and existing relative treatment effect information, the matching-adjusted indirect comparison estimate can be obtained, as well as confidence intervals and standard errors.

To conclude, further development of the R package is recommended to encourage consistency in the MAIC implementation such that the technique may be applied in line with established best practice, although this itself is an area in need of further research.

Acknowledgements

This research was made possible through the Employment Based Postgraduate Scholarship from the Irish Research Council. Under the terms of this grant, I conducted this research under the dual status as Research Master's student of the Higher Education Institution, Trinity College Dublin and employee of the employment partner, Novartis Ireland. The collaboration was made possible through a supervisory team: the principal investigator and primary supervisor was Dr. Arthur White, the secondary supervisor was Dr. James O'Mahony and the employment mentor was Dr. Ronan Mahon.

Owing to the unique partnership characterising this Research Master's, there are a number of people to acknowledge for making it happen. I wish to personally thank the Irish Research Council for supporting the funding of this Master's and giving me the opportunity to conduct research in this very special discipline that I have thoroughly enjoyed working within every day. I hope their excellent funding initiative continues to benefit researchers in Ireland for many years to come. I also wish to thank all the staff involved from both Trinity College Dublin and Novartis Ireland who worked behind the scenes to put the necessary arrangements in place to facilitate this research.

I would like to thank my supervisor, Dr. Arthur White, who not only put in place the necessary steps for this Research Master's to happen but provided invaluable support throughout. He helped to facilitate me presenting my work at my first conference, for which I am extremely grateful, and he is never short on ideas! To Dr. James O'Mahony, who kindly agreed to act as secondary supervisor, I thank him for his continued mentorship, his proofreading, as well as bringing his vast experience in academia and health economics to the team, which complemented the project. For bringing on board the participation of Novartis Ireland and being a considerate and approachable employment mentor, I thank Dr. Ronan Mahon. I feel very fortunate to have received his guidance inspired from his many varied academic and industry experiences, as well as the cúpla focal, go raibh maith agat!

To all the other health economists in Novartis Ireland that I had the pleasure of working with over the course of this Research Master's, I would like to thank them for introducing me to their work in applying health economics in an industry setting. I am in awe at the level and sheer breadth of skill required in the health economic modelling team!

For their help in getting me started on this project, I wish to thank Dr. Joy Leahy and Dr. Howard Thom. Their passion for this discipline made them excellent sources of advice and suggestions on progressing the project.

Thank you to my fellow postgraduate researchers in the statistics department for reminding me of the importance of the academic setting for this project. I am glad to have gained their insights

during my first experience of postgraduate research and most especially at my first conference, CASI. I also wish to thank Yi Shu Lin for her help in proofreading.

I would finally like to acknowledge a former academic staff member at the statistics department in Trinity College and former lecturer of mine, Prof. Brett Houlding, who sadly passed away shortly before I completed this thesis. Although not directly involved in this work, he taught many of my undergraduate courses in statistics covering a wide variety of subjects, and was tremendously helpful in this regard. I would like to thank him for his dedication to the teaching and for the expertise he provided to me and my fellow classmates.

Contents

LIST OF TABLES	VIII
LIST OF FIGURES	VIII
GLOSSARY	IX
CHAPTER 1 INTRODUCTION	1
1.1 THESIS OVERVIEW	2
1.2 OVERVIEW OF CHAPTERS	2
CHAPTER 2 BACKGROUND	4
2.1 EVIDENCE SYNTHESIS CONCEPTS	4
2.1.1 META-ANALYSIS	5
2.1.2 INDIRECT TREATMENT COMPARISONS AND THE BUCHER METHOD	7
2.1.3 MIXED TREATMENT COMPARISONS	8
2.1.4 SCALE TRANSFORMATION	10
2.1.5 FURTHER CHALLENGES IN EVIDENCE SYNTHESIS	12
2.2 THE CONTEXT FOR POPULATION ADJUSTMENT METHODS	13
2.3 BACKGROUND TO EPIDEMIOLOGY AND TREATMENT OUTCOMES	14
2.3.1 BIAS, CONFOUNDING AND EFFECT MODIFICATION	15
2.4 LOGISTIC REGRESSION	16
2.4.1 THE EXPONENTIAL FAMILY OF DISTRIBUTIONS	16
CHAPTER 3 LITERATURE REVIEW	18
3.1 BASIS FOR POPULATION ADJUSTMENT METHODS	18
3.1.1 PROPENSITY SCORE METHODS	19
3.1.2 CALIBRATION	19
3.1.3 INTRODUCTION TO MAIC	20
3.2 SIMULATED TREATMENT COMPARISON (STC)	23
3.3 COMPARISON OF POPULATION ADJUSTMENT METHODS	23
3.4 REVIEW METHODOLOGY	25
3.5 MAIC METHODOLOGY	26
3.5.1 RELATING THE LOGISTIC MODEL WITH RELATIVE TREATMENT EFFECT	31
3.5.2 VARIANTS OF MAIC METHODOLOGY	32
3.5.3 KEY ASSUMPTIONS UNDERLYING THE METHODS	34
3.5.4 RECOMMENDATIONS FROM TECHNICAL SUPPORT DOCUMENT 18	35
3.6 REVIEW OF MAIC APPLICATIONS	36

3.6.1	OVERVIEW OF THE LITERATURE	37
3.6.2	FINDINGS FROM SELECTED MAIC APPLICATIONS	39
3.6.3	SUMMARY OF REVIEW OF MAIC APPLICATIONS	44
3.7	EXISTING R PACKAGES	45
3.8	OVERALL LITERATURE REVIEW SUMMARY	46
CHAPTER 4 R PACKAGE		48
4.1	MOTIVATION	48
4.2	PACKAGE DEVELOPMENT WORKFLOW	49
4.3	REQUIREMENTS	51
4.4	DEVELOPMENT	52
4.4.1	OBJECT ORIENTATION	53
4.4.2	CHALLENGES IN MATCHING TO A MEDIAN STATISTIC	54
4.4.3	DESIGN	55
4.5	PACKAGE WALKTHROUGH	59
CHAPTER 5 CONCLUSIONS		69
5.1	FURTHER WORK	69
5.2	FINAL REMARKS	71
APPENDIX A BINOMIAL DISTRIBUTION AS A MEMBER OF THE EXPONENTIAL FAMILY OF DISTRIBUTIONS		72
APPENDIX B LIST OF REVIEWED MAIC PUBLICATIONS		74
BIBLIOGRAPHY		75

List of Tables

TABLE 1 OUTCOMES FOR BINARY DATA TABULATED	6
TABLE 2 ECOLOGICAL BIAS	13
TABLE 3 THE SEVEN COMMON COMPONENTS TO A STANDARD R PACKAGE.....	49
TABLE 4 SUMMARY OF OBJECTS AND FUNCTIONS IN R PACKAGE.....	57

List of Figures

FIGURE 1 DEFINITION OF PROGNOSTIC VARIABLES AND EFFECT MODIFIERS	5
FIGURE 2 INDIRECT COMPARISON PERFORMED BETWEEN A AND B IN ADDITION TO A DIRECT COMPARISON	8
FIGURE 3 EXAMPLE OF A NETWORK OF TREATMENTS WITH EXAMPLE INDIRECT COMPARISONS INDICATED BY DASHED LINES	9
FIGURE 4 ANCHORED INDIRECT COMPARISON INDICATED BY DASHED LINE.....	21
FIGURE 5 UNANCHORED INDIRECT COMPARISON INDICATED BY DASHED LINE.....	22

Glossary

AgD	Aggregate Data
EUNetHTA	European Network for Health Technology Assessment
AIC	Akaike Information Criterion
ASAS	Assessment of SpondyloArthritis international Society
BFGS	Broyden, Fletcher, Goldfarb and Shanno
CADTH	Canadian Agency for Drugs and Technologies in Health
CMD	Command
CRAN	Comprehensive R Archive Network
DIC	Deviance Information Criterion
DSU	Decision Support Unit
ESS	Effective Sample Size
GLM	Generalised Linear Model
HTA	Health Technology Assessment
IPD	Individual(-level) Patient Data
ISPOR	International Society for Pharmacoeconomics and Outcomes Research
ITC	Indirect Treatment Comparison
MAIC	Matching-Adjusted Indirect Comparison
MLE	Maximum Likelihood Estimation
MTC	Mixed Treatment Comparison
NCPE	National Centre for Pharmacoeconomics Ireland
NICE	National Institute for Health and Care Excellence
NMA	Network Meta-Analysis
OR	Odds Ratio
OS	Overall Survival
PATE	Population Average Treatment Effect
PFS	Progression Free Survival
RC	Reference Classes
RCT	Randomised Controlled Trial
SATE	Sample Average Treatment Effect
SMDM	Society for Medical Decision Making
STC	Simulated Treatment Comparison
TSD	Technical Support Document

Chapter 1 Introduction

Currently, all new pharmaceutical products proposed for reimbursement by the state healthcare provider in Ireland (and in other countries subject to their own regulatory bodies), must be reviewed by the National Centre for Pharmacoeconomics Ireland (NCPE) via a Health Technology Assessment (HTA). This assessment will typically involve a cost-effectiveness analysis and in order to establish the effectiveness of the product in relation to relevant alternatives, comparative effectiveness research (often referred to as relative effectiveness in Europe) is carried out. The aim of this thesis is to establish and articulate an understanding of population-adjusted treatment comparison methods and in particular, assess the potential impact of Matching-Adjusted Indirect Comparisons (MAIC) on HTA.

For new pharmaceutical products and medical treatments, there is often a small amount of data to work with in terms of assessing the product/treatment for its effect on patient outcomes and/or its overall true cost. Ideally, new treatments would be tested in randomised trials against all relevant alternatives; typically, a drug is assessed against what is known as the current *standard of care*. However, in the case of new treatments, it may take several years before a direct study is performed that compares the treatment with other viable alternatives available at the time. This delay hinders healthcare decision-makers in approving such new treatments for reimbursement.

This research will examine a technique that may be used to inform healthcare decision-makers of the relative effectiveness and hence modelled cost-effectiveness of new treatments being proposed for use in Ireland or elsewhere. The technique is referred to as Matching-Adjusted Indirect Comparisons. MAIC is a relatively novel technique that was first published in 2010 (Signorovitch et al. 2010). The full implications of employing this technique remain poorly understood.

MAIC can compare different treatment outcomes among separate trials in accordance with one of two approaches. The first of these makes use of clinical trial data on a common comparator (such as placebo). This comparison is referred to as *anchored*. The other approach makes the comparison without data on a common reference arm. This comparison is referred to as *unanchored*.

There are many reasons for why there may be no data available on a common comparator treatment against two competing interventions. This is primarily driven by forces that dictate clinical trial planning. Clinical trials can be costly for investigators to run and typically demand significant lengths of time in the course of their planning, execution and analysis. Therefore, the comparators that have been assessed directly against the intervention may have been selected according to criteria that are not necessarily linked to the latest decision question. In the course of running a lengthy clinical trial, the common comparators identified as relevant may change. In some cases,

however, it is unethical to randomly assign patients to a particular treatment arm of a clinical trial where clinical equipoise cannot be assured. Clinical equipoise refers to the principle that there is no pre-existing evidence that one treatment is superior to another. Hence, where previous evidence may exist that suggests that a treatment has some benefit, a single-arm clinical trial may be conducted where all patients are instead assigned to a new treatment and observed. This can often occur where patients have poor prognosis and whose treatment options are limited, particularly in the case of oncology.

The MAIC method uses individual patient data (IPD) to equalise baseline characteristics across trials from different studies. In theory, this will reduce potential biases and ensure that fairer comparisons can be made. Since the full implications that may arise from use of this technique are not fully known, in assessing its potential impact, the aim is to identify best practice for its use in HTA according to a standardised implementation. This is achieved through modelling in the R programming language.

1.1 Thesis Overview

This thesis will investigate the contribution of MAIC to comparative effectiveness research for the purposes of HTA. In order to do so, some background material is presented to establish the framework of evidence synthesis. This is concerned with how the total body of evidence available (in its various forms) may be combined together to give the best picture of the true effectiveness of a pharmaceutical product or medical treatment. This draws upon theory from both epidemiology and statistics.

Following this, the MAIC technique, based on past publications, is established through means of a literature review. In addition to the literature surrounding the MAIC methodology, a review of existing MAIC applications to pharmaceutical products for the purposes of ascertaining relative treatment effect is carried out.

Some insights from the review are presented and lead to the development of a package in the R programming language (R Core Team 2018). The aim of the R package is to make implementation of the technique easily accessible, reproducible and thereby convert the existing theory on the MAIC method into a tool of use to industry and the health economics research community. The package contains a number of functionalities but can be extended further. Some suggestions for doing so are provided and the overall contribution of this thesis is summarised.

1.2 Overview of Chapters

There follows an overview of the ensuing chapters of this thesis:

- Chapter 2 Background – Some concepts and existing practices encompassed within evidence synthesis for the purposes of comparative effectiveness research are set out in order to provide a foundation for the central focus of this thesis. Epidemiology and statistics

have a crucial role to this framework and some additional concepts within these fields are described in this chapter.

- Chapter 3 Literature Review – This chapter details a comprehensive literature review of three parts aimed at establishing the current level of existing understanding of the MAIC technique. The first examines the methodological papers for MAIC and another method for population adjustment to determine the commonly accepted definition of the MAIC method. The second part examines numerous comparative effectiveness publications that apply the method to pharmaceutical products, noting use of method variations. The final part of the review explores existing R packages of relevance to the method.
- Chapter 4 R Package – The process of developing a package in the R programming language is described in Chapter 4. Details include the rationale for developing R packages, the requirements for the new package, the process of development and an overview of the final design.
- Chapter 5 Conclusions – The final chapter summarises the work described by this thesis, in particular, consolidating understanding of the MAIC technique as used in HTA, the future use of MAIC and some suggestions for further work in this area in terms of extending the R package are also presented.

Chapter 2 Background

This chapter introduces a number of epidemiological and statistical concepts relevant to both MAIC and more broadly, comparative effectiveness research for HTA. This begins with various models used for evidence synthesis including meta-analysis. This is followed by an exploration of the complications that can arise from combining different measures of study effect from many studies in relation to scale. Issues pertaining to indirect comparisons, the use of aggregate summary information and heterogeneity are identified. The idea of population adjustment is then introduced followed by some epidemiological concepts and the basis for logistic regression, which is instrumental to MAIC.

2.1 Evidence Synthesis Concepts

In order to establish evidence-based healthcare decision-making, all relevant competing interventions for a given condition need to be considered (Jansen et al. 2011). Relevant competing interventions may have already been compared together in a randomised controlled trial (RCT) and can therefore be directly compared for effectiveness (a direct comparison). This is assuming the same outcome type was recorded for both treatments as would be standard practice in a clinical trial setting.

Randomisation protects against selection bias, which arises when, before being administered treatment, there are underlying systematic differences between the groups to be compared and these affect response to a therapy. Responses will be recorded for each patient. Then, for each treatment, an average treatment effect can be obtained, specifically, a Sample Average Treatment Effect (SATE), which is not to be confused with the true Population Average Treatment Effect (PATE). This would be obtained if the treatment was administered to the entire population versus the control and would be the ultimate quantity we would wish to obtain. Administering to the entire population is unlikely to be practicable. Taking the difference between two treatments in terms of (sample) average treatment effect will return the relative treatment effect and this is the key measure for ascertaining the most effective treatment (Eichler et al. 2010, Phillippo et al. 2016). There may exist several RCTs which have been conducted independently and compare the same two treatments, each potentially returning a different estimate of relative treatment effect.

We can consider the observed summary outcome for a treatment arm of a randomised controlled trial to be made up of three parts:

- The effect owing to the treatment itself
- The effect as a result of patient characteristics
- The effect as a result of study characteristics

If a randomised controlled trial is placebo-controlled, the placebo arm is considered to identify the effect owing to the study and patient characteristics. The study and patient characteristics

contributing to the outcome expressed within the placebo arm are referred to as *prognostic variables*; they have an impact on the absolute outcomes for both the intervention and placebo arm. Correspondingly, the summary outcome recorded in the intervention arm is considered to include the effect owing to the treatment itself in addition to that owing to the study and patient characteristics. Prior to treatment administration, patients are randomised across the two arms such that known and unknown prognostic variables are balanced. The individual effect each prognostic variable contributes in the placebo arm is then also present in the intervention arm.

Meanwhile, study and patient characteristics which have an impact on the difference between the intervention and placebo arm are known as *effect modifiers*. This means that for particular subgroups, the treatment effect in the intervention may appear to vary. Effect modifiers are not necessarily also prognostic variables but where an effect modifier is also a prognostic variable, then the outcome in the placebo arm is associated with the treatment effect. Figure 1 summarises these key definitions.

- Prognostic variable – as the value of a covariate varies, absolute outcomes are affected, but to the same extent for all compared treatments.
- Effect modifier – as the value of a covariate varies, the effect of the treatment is altered. Absolute outcomes vary among subgroups of patients. If this pattern is different between treatments, then the relative treatment effect is impacted. An effect modifier is not necessarily also a prognostic variable.

Figure 1 Definition of prognostic variables and effect modifiers

These distinctions become more crucial when comparing across multiple trials and multiple treatments. The methodologies for doing so are described in the ensuing subsections. It is worth noting that when multiple trials are characterised by differing levels of patient characteristics, it is said that heterogeneity exists between the trials or the trials are heterogeneous.

2.1.1 Meta-Analysis

In order to utilise all the available evidence derived from numerous studies, the multiple estimates of relative treatment effect can be combined together in what is known as a meta-analysis (Higgins and Green 2011). This combines the results of the different RCTs by calculating a summary statistic for each before calculating a weighted average. The general concept of combining results together from multiple sources is often referred to as *evidence synthesis*. The summary statistic used may be in the form of an odds ratio, risk ratio, risk difference, hazard ratio, difference in means or standardised mean difference. This manner of pooling the effect size can be achieved through one of two approaches, a fixed effects model or the random effects model (Hoaglin et al. 2011).

The first of these assumes that there is no heterogeneity between the study populations and that the true treatment effect being estimated in each study is the same. Any variability arising from

between studies is entirely due to random variation owing to sampling error. That is, the trial population did not capture the behaviour of the full population and true effect. Smaller sample sizes are inherently at more risk of such error. In calculating the weighted average, studies which have greater precision are assigned a greater weight. Precision is simply assumed to be greater for studies that have larger numbers of participants. Therefore, under the inverse variance method in the case of continuous outcome data, each i^{th} study of n studies has an effect estimate, \hat{d}_i that is weighted by the inverse of the variance in its estimate: $w_i = \frac{1}{\sigma_i^2}$. The fixed effects model assumes each estimate will yield an overall estimate for the treatment effect as: $\hat{d} = \frac{\sum_{i=1}^n w_i \hat{d}_i}{\sum_{i=1}^n w_i}$.

If the studies have recorded binary outcome data and report event rate data such as the odds ratio, the relative risk or the risk difference, then the Mantel-Haenszel method is used to weight studies. This entails weights calculated using the values from a typical outcome table depicted as follows:

	Event	Non-Event
Treatment	a	b
Control	c	d

Table 1 Outcomes for binary data tabulated

The odds ratio is calculated as: $\frac{a/b}{c/d} = \frac{ad}{bc}$ and the weight is defined as: $w_i = \frac{bc}{n}$.

The risk ratio is calculated as: $\frac{a/(a+b)}{c/(c+d)}$ and the weight is defined as: $w_i = \frac{(a+b)c}{n}$.

The risk difference is calculated as: $\frac{a}{a+b} - \frac{c}{c+d}$ and the weight is defined as: $w_i = \frac{(a+b)(c+d)}{n}$.

In the random effects model, it is assumed that each study is being performed on a range of populations. Therefore, each study is estimating one of a range of true treatment effects. These are assumed to be normally distributed $\delta_i \sim N(\mu, \Delta^2)$ with a common mean μ and this is the value of interest.

To summarise, the fixed effects model describes each treatment effect observed from each study as follows: $\hat{d}_i = d + \epsilon_i$ whereas in a random effects model (as proposed by (DerSimonian and Laird 1986)), each study treatment effect, \hat{d}_i , is assumed to be itself an estimate of its own unique treatment effect, δ_i . In contrast to the fixed effects model, the true treatment effect is described as a distribution with mean μ . The random effects model therefore describes each observed study treatment effect as follows: $\hat{d}_i = \mu + \delta_i + \epsilon_i$ (Hoaglin et al. 2011).

This model is typically more useful than the fixed effects model as homogeneity within study populations would be a difficult condition to satisfy, but it yields less precise results, particularly where between-trial heterogeneity is large. The model choice is not straightforward since a random

effects model takes greater consideration of smaller studies which could themselves be returning biased results. Some measures have been defined for assessing heterogeneity and are detailed in (Higgins and Green 2011) among many sources.

The study effect results from the meta-analysis may be summarised and graphed together in a forest plot. However, a combined estimate may be biased due to heterogeneity between trials i.e. the underlying differences in trial design, methodologies or study populations. This entire framework as outlined above and described in the following sections has been well summarised in a guideline document produced by the European Network for Health Technology Assessment (EUNetHTA 2013).

2.1.2 Indirect Treatment Comparisons and the Bucher Method

A separate issue that frequently arises in comparing available treatments is that new treatments may not be directly compared with the most relevant treatment options for the given indication. On the other hand, there may be a common comparator against which both options were examined in clinical trials. Using the common comparator to compare two treatments is referred to as an indirect treatment comparison (ITC) (in this thesis, indirect comparison is used). A method for performing such a comparison was first proposed by (Bucher et al. 1997). The authors highlight that in the past, researchers would pool results from only the active treatment arms of RCTs and make a comparison across these trials. This “naïve unadjusted comparison” is no better than working with observational study data (where patients are merely observed and not randomised) since the patients are all drawn from different, independent cohorts instead of being drawn randomly from the same pool. This allows for other factors, aside from the treatment that was administered, to significantly impact the patients’ response and thus bias the comparison.

In a naïve comparison, all prognostic variables would have to be accounted for as they are likely to be imbalanced between the two trials. This is a substantial requirement which would in most cases be impossible to achieve. Naïve comparisons are therefore seldom accepted in the context of health technology assessments or comparative effectiveness research.

Instead, Bucher et al. (1997) proposed an adjusted method which makes use of the control arm in each RCT (in effect, a pre-existing direct comparison within the trial data). For a trial with a binary outcome, assuming the odds ratio as the measure of treatment effect and direct evidence available for B vs A and C vs A (where necessary each comparison may be derived as a weighted average of multiple study estimates), the indirect comparison is:

$$\ln(OR_{indirect\ BC}) = \ln(OR_{BA}) - \ln(OR_{AC})$$

This assumes that among the studies, variables which have an impact on the treatment effect for particular subgroups behave in the same way for both treatments in each RCT. In other words, effect modifiers must be balanced. Owing to the randomisation taking place within trials B vs A and

C vs A, it is possible to make a fair comparison of B and C even if the levels of the prognostic variables are different. This is further described in (Phillippo et al. 2016) in terms of *effect modifiers* and the assumption of *constancy of relative effects*. Bucher et al. (1997) theorised the approach for odds ratios but the same theory may be implemented for other measures, as documented by the Canadian Agency for Drugs and Technologies in Health (CADTH), the HTA body in Canada (Wells et al. 2009). This is also described in Sections 2.1.4 and 3.2.

2.1.3 Mixed Treatment Comparisons

Even when a direct comparison is possible between two treatments, an indirect comparison may still be obtained. For example a closed 'loop' of three treatments (A, B, C), where all treatments have been compared against each other directly could be informed by what is known as a multi-arm trial; where a single cohort of patients are randomly assigned to more than two groups with a different treatment administered in each. A three-arm trial comparing A, B and C can be very useful, however Treatment A might additionally be compared to Treatment B via common comparator C. Here a separate estimate would be obtained. Combining this direct evidence with the indirect evidence is referred to as a mixed treatment comparison and this can improve the precision of the estimate (Jansen et al. 2011). This scenario can be visually represented in a simple network diagram and is depicted in Figure 2 below. Here, the nodes represent treatments and each solid connecting edge represents a typical two-armed RCT between treatments. Unfortunately, one limitation of this schematic is in distinguishing between a three-arm trial and a set of three two-arm trials. One way to include this information would be to simply label each edge with the trial name or possibly introduce a colour coding system for more complex networks. An edge displayed as a dashed line represents indirect evidence that may be generated as a result of two or more pieces of connected direct evidence.

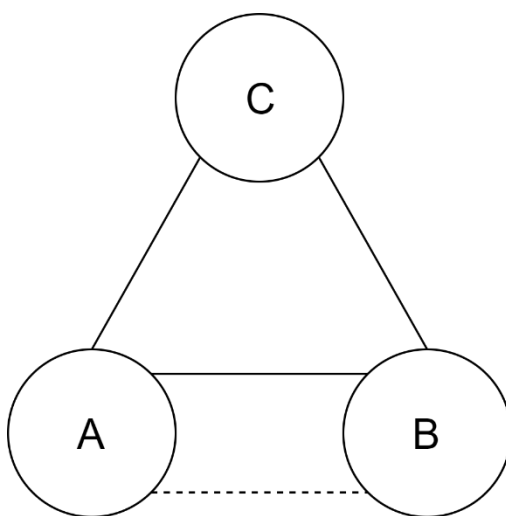


Figure 2 Indirect comparison performed between A and B in addition to a direct comparison

Multiple treatments may be linked together directly and/or indirectly through what is known as a network meta-analysis (NMA) (Lumley 2002, Sutton et al. 2008). An example of this is illustrated in Figure 3:

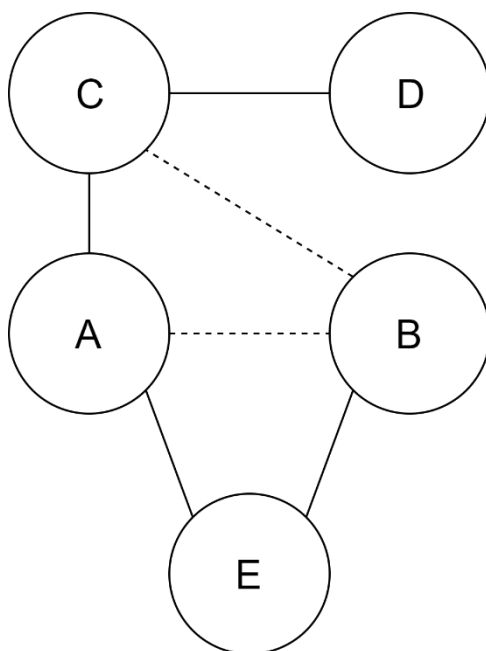


Figure 3 Example of a network of treatments with example indirect comparisons indicated by dashed lines

As can be observed in the figure above, a treatment may be linked through a 'chain' of common comparators extending the network further, but this increasing order of comparison tends to increase the estimate's standard error (Hawkins et al. 2009).

It is important to note that mixed treatment comparisons rely on the differences in the distribution of effect modifiers to be the same on average between the set of trials making comparisons between B vs A and the set of trials comparing C vs A. If the levels of effect modifiers vary between trials in each set but the pattern of variation is the same i.e. on average they are the same between the two sets, a random-effects model can be used to account for these differences to avoid biased estimates from a mixed treatment comparison. It is worth noting that the requirement to account for imbalanced levels in effect modifiers between trials applies to all effect modifiers, irrespective of whether they have been measured in the trials or not. It must therefore be acknowledged that there is always a risk of some element of bias in a mixed treatment comparison even when observed effect modifiers are balanced.

This leads to an important point regarding the applicability of the results from indirect comparisons generally. Such comparisons depend on the trial populations involved in the treatment comparison to be the same, or rather, that underlying population characteristics that have an impact on the treatment comparison are accounted for, such that population differences are in effect eliminated.

The results from such indirect comparisons can only be considered for the mixed population defined over the included trials. While robust systematic reviews will attempt to capture all eligible studies, it is possible that the results from an analysis involving a select number of trials is only applicable to the merged population and not generalisable to a wider population of interest.

This can be a significant challenge in the course of health technology assessment if the objective is to determine whether a new health technology represents a meaningful benefit for a wider patient population that would be treated within the public health system. For this reason, some heterogeneity in the scope of patient populations across trials may increase the external validity of the analysis, provided that there is little heterogeneity in the distribution of effect modifiers across studies (Jansen et al. 2011). The issue of applicability of results to specific populations is relevant in the discussion of methods for population-adjusted treatment comparisons, since this differs from indirect comparison methods, as described in this thesis.

2.1.4 Scale Transformation

Methodology for pairwise meta-analysis is well established, however, it becomes more challenging to combine evidence from multiple data sources including direct, indirect and mixed treatment comparisons, network meta-analysis and multi-arm trials. There may be a great variety of outcome types reported in trials from this body of evidence.

A requirement for the indirect comparison, including MAIC, is that the two components being compared are measured on a scale where the treatment effects are additive. This means that it is true to say that the sum of the two components does in fact result in the numerical addition of the two values. This relationship between the two components may not hold if measured on a different scale; the relationship may be multiplicative for example, or nonlinear.

The scale on which the outcomes are observed may be referred to as the natural outcome scale. This may not be the same as the linear predictor scale; where treatment effects, effect modification and prognostic variables are additive. Indirect comparisons require outcomes to be recorded on the linear predictor scale. In order to achieve linearity, data transformations are required in the form of a link function $g()$ and this has become standard practice for performing an indirect comparison (Dias et al. 2013, Dias et al. 2011b). A common example is the log transformation for the binary outcome onto the log odds scale where the natural outcome scale for binary outcomes is the probability scale.

For binary outcome data consisting of k arms of trial i , events r_{ik} are expressed as follows:

$$r_{ik} \sim \text{Binomial}(p_{ik}, n_{ik})$$

Where p_{ik} is the probability of an event in arm k of trial i .

A typical linear model is inappropriate for a binary outcome. A simple inspection of event occurrence against treatment group will illustrate that the dependent variable is not normally distributed. Using a linear regression model would predict values for the event occurrence ranging on the set of real numbers \mathbb{R} and this is clearly not useful. However, the odds scale provides a useful range on which to map the probability of an event occurring. This is because the range is no longer restricted to the $[0,1]$ interval as is the case with probability. The odds are instead written as $p:1-p$ or $\frac{p}{1-p}$ and this fraction tends to ∞ as $p \rightarrow 1$. Taking the log of the odds; $\log\left(\frac{p}{1-p}\right)$ will map to a continuous range from $-\infty$ to ∞ and so modelling on this scale would potentially offer a good fit for binary outcome data (Hosmer and Lemeshow 2000). Section 2.4.1 describes further how the link function, the logarithm of odds or logit is related to the binomial distribution.

Therefore, for a fixed effects model, the probability of experiencing an event for a patient in trial i is modelled as:

$$\text{logit}(p_{i1}) = \mu_i$$

$$\text{logit}(p_{i2}) = \mu_i + d$$

Where μ_i represents the log odds of the outcome for the control arm of trial i and d is the log odds ratio of achieving the event. For a random effects model, the equation for the treatment arm is instead:

$$\text{logit}(p_{i2}) = \mu_i + \delta_{i2}$$

Where δ_{i2} is a trial-specific estimate of treatment effect and is distributed according to: $\delta_{i2} \sim N(d, \sigma^2)$. This is further described in Section 3.5.1 through a published example implementation.

2.1.5 Further Challenges in Evidence Synthesis

There are, however, problems associated with indirect comparisons. Patients are randomised in an RCT but this randomisation property does not hold across different RCTs (Bucher et al. 1997, Jansen et al. 2011). As noted in section 2.1 in respect of effect modifiers, prognostic variables and meta-analysis; conditions may easily differ between RCTs, such as the geographic setting in which the study was performed or differing patient characteristics, for instance, the average patient age. If these characteristics are effect modifiers and hence have an impact on relative treatment effect, then the collection of studies is said to be heterogeneous (Jansen et al. 2011). For this reason, simply comparing the (sample) average treatment effect for each treatment obtained in separate clinical trials is insufficient. Indirect comparisons will not be able to account for differences in relative treatment effect. On the other hand, differences in purely prognostic variables, which affect absolute outcomes only, may be accounted for in an indirect comparison. Once a common comparator is available between the two treatments being compared and both treatments have been randomised against the common comparator in separate clinical trials, the portion of the observed effect that can be attributed to these prognostic variables will not contribute to the indirect treatment comparison. Determining whether differing patient characteristics between trials have an impact on solely absolute or also relative treatment effects represents a significant challenge (Sutton et al. 2008) and is considered in Section 3.5.

For robust decision-making, evidence on the relative effectiveness between competing treatment options should come from studies that are comparable (Higgins and Green 2011, Jansen et al. 2011). That is, the characteristics of the studies should be as similar as possible. Conversely, if for example, disease severity is greater in trials comparing drug A with B than trials comparing A with C, and differences such as these have an impact on relative treatment effect, then the estimate of the indirect comparison, B with C, may well be biased (Jansen et al. 2011, Signorovitch et al. 2010).

The treatment comparisons made within this framework are based on published clinical trials which report summary statistics quantifying the treatment effect. This could be reduction in systolic blood pressure, for example. These treatment outcomes are typically reported as aggregate data (AgD). Further detail on the various forms these outcomes may take is presented in Section 2.3. We may be able to tell from this summary data how similar two trials are in terms of patient characteristics and the study literature should indicate the various conditions of the trial. It may then become apparent that the two studies are far from comparable and any indirect comparison made may be undermined by bias.

It is also recognised that comparing aggregate data alone is prone to ecological bias. This refers to the case where an initial comparison of AgD has led to one relative effect but further inspection of data at the individual level has resulted in a different estimate in relative effect. This is due to associations present at the individual level not exhibited at the summary level (Greenland and

Morgenstern 1989). This can lead to an entirely opposing conclusion about the effectiveness of one drug versus another. To illustrate, hypothetical patient outcomes are presented in Table 2 below. On inspection of the AgD, treatment C appears more effective, stratifying by baseline severity of disease shows that treatment B is in fact more effective:

	A vs B trial (80% Severe cases)		A vs C trial (20% Severe Cases)		B vs C indirect comparison
Patient Outcome (% response)	Treatment A	Treatment B	Treatment C	Treatment A	$(B - A) - (C - A)$
AgD	10	46	54	16	-2
<i>Stratifying Patient Outcomes by Baseline Severity of Disease</i>					
Severe Cases	8	40	38	8	2
Non-severe Cases	17	70	58	17	12

Table 2 Ecological Bias

It can be concluded that, where available, the use of IPD in the analysis of clinical trial data is more preferable when establishing the treatment effect of a drug, due to the greater amount of information available describing the distribution of patient covariates and outcomes. As this section has shown, the sheer volume of evidence can become overwhelming, particularly when taking into account the different types of evidence that may be derived through various methods as well as the numerous combinations within. Methodologies for evidence synthesis continue to represent an active area of research, featuring prominently at SMDM and ISPOR conferences. There are centres of research dedicated to this topic including the Johns Hopkins Center for Clinical Trials and Evidence Synthesis, Evidence Synthesis Ireland and Cochrane. This thesis will focus on the role of methods of population adjustment within this framework.

2.2 The Context for Population Adjustment Methods

To summarise, there are situations where making an indirect comparison is highly desirable. However, the potential for heterogeneity inherent in separate trial populations can make estimates arising from indirect comparison less likely to reflect true comparative effectiveness. If other sources of heterogeneity in the trials can be eliminated and only the differences in the populations of the two trials remain, it may be possible to adjust the population characteristics of one trial using the individual-level patient data.

An extensive document detailing these methods and their preferred usage in health technology submissions to the respective HTA body in the United Kingdom, is Technical Support Document 18 (Phillippo et al. 2016). This was produced by the Decision Support Unit (DSU): a collaboration between several universities in the UK which is funded by the HTA agency, the National Institute for Health and Care Excellence (NICE). The DSU is commissioned to provide research capacity in

support of Technology Appraisal (a term typically used by NICE instead of HTA), normally as a result of identification of areas of interest by appraisal committees, academic groups or NICE staff. The Technical Support Documents (TSDs) are intended as guides for implementing specific methods so as to inform interested stakeholders such as pharmaceutical companies and assessment groups on how to approach incorporating these into their HTAs.

In the case of TSD 18, it had been noted that MAICs were being applied in numerous technology appraisals with little to no understanding of the resulting implications (Phillippo et al. 2016). Therefore, the document was developed by members of the group at the University of Bristol and since its release has gone a long way towards consolidating knowledge surrounding these methods and acting as a key point of reference. Where relevant, the contributions made by the TSD are referenced throughout Section 3.5 and a summary of the authors' recommendations with respect to HTA submissions to NICE are included in Section 3.5.4.

2.3 Background to Epidemiology and Treatment Outcomes

In order to achieve the goal of best decision-making using evidence-based medicine, measurement of treatment outcome becomes particularly important if the most effective treatments are to be identified. Cochrane, a global network organisation made up of voluntary experts who aim to promote evidence-based health decision-making, list the various types of outcome data they recognise to be used in clinical studies in their Handbook for Systematic Reviews of Interventions (Higgins and Green 2011). These are: dichotomous, continuous, ordinal, time-to-event (survival) and counts & rates.

For each of these data types, more than one effect measure may be possible. For dichotomous data, these include risk ratio, odds ratio and risk difference. For continuous data, mean difference or standardised mean difference may be used. Hazard ratios may be used in time-to-event data. Rate ratios may be used for counts and rate data. Ordinal data may be described in terms of proportional odds ratios depending on the number of categories involved (Higgins and Green 2011). The issue of this choice of scale on which to measure the effect is repeatedly cited in TSD 18 (Phillippo et al. 2016) and it has been noted that the choice of effect measure can have an impact on ranking treatments for probability of being most effective in a mixed treatment comparison (Norton et al. 2012).

In this thesis, we will focus on the case where the patient outcome is binary, referring to an outcome taking values $x \in \{0,1\}$ (as a matter of distinction, dichotomous can be generalised to a variable taking any two distinct values and not exclusively $\{0,1\}$). It is important to first determine whether the outcome being recorded is of positive benefit to the patient or represents a negative outcome, such as experiencing a cardiovascular event. This will affect the model and interpretation of the results. For a trial of an experimental treatment with a control arm, events can be recorded as in Table 1 and an odds ratio may be calculated. When using the odds ratio effect measure for an

indirect comparison, a transformation on to the log odds ratio scale may be required. In plotting or tabulating results comparing various treatments, it is important to determine whether a negative log odds ratio i.e. lower odds of experiencing the event, favours or goes against the treatment.

2.3.1 Bias, Confounding and Effect Modification

Bias is an important concept in epidemiology and statistics. In the latter, it refers to results deviating from the truth due to a systematic error. In other words, it would not be possible for the estimates to reflect the true value because even before they are generated, the methodology adopted for doing so confines them to a set of values that are misaligned from the true value. Some key distinctions in the terminology are outlined below:

- A reliable estimate is often described as one which is reproducible; the returned values are consistent after calculating the estimate multiple times. However, a more distinguishing term to describe this property could be 'consistent'. The term 'reliable' is problematic as it often carries different meanings across different disciplines and among individuals. More specific terminology follows.
- A precise estimate is one where the variation between iterations of the estimate owing to random error is small. This is a more specific and hence preferable term to 'reliable'. A reliable estimate could (reliably) produce the same set of values that are disparate and distributed in equal proportions over multiple simulations.
- An accurate estimate is one which captures the true value. An estimate which is inaccurate could also be described as 'invalid'.

A biased estimate may be reliable, precise and without random error but will be fundamentally invalid. In contrast, a clinical trial conducted without systematic error can still be subject to random error and imprecision, but many replications of the same study should produce results averaging to the true value. Systematic error can be caused by one of many possible factors. In epidemiology, the types of bias typically referred to are as a result of some aspect of the study design or analysis of the data. This could be the way in which participants of a study are recruited, selected, retained or the way in which their characteristics and responses are measured (Higgins and Green 2011).

Confounding refers to the situation where an association between a treatment and outcome exists but is distorted by another variable that is not the causal factor itself (Greenland and Morgenstern 1989). Randomisation is used to minimise potential distortions arising from confounding variables. One way to reveal the actual association is to stratify the data into various subgroups of the confounding variable, as previously discussed in relation to ecological bias in Section 2.1.5.

Effect Modification refers to the case where a variable has an effect on the outcome, but the degree of impact varies for different subgroups of another variable. This is often referred to as interaction in statistics. Effect modification more strictly relates to a biological phenomenon observed in the

behaviour of a disease, as opposed to an association observed simply from data (Kamangar 2012). In practice, the term is often used interchangeably with interaction.

2.4 Logistic Regression

Logistic Regression is important to the MAIC method for both generating weights and calculating relative treatment effect. It is also central to the other method of population-adjusted treatment comparisons; the Simulated Treatment Comparison.

2.4.1 The Exponential Family of Distributions

Continuous data may be analysed via a linear model which can be represented as follows:

$$E(Y_i) = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

$$Y_i \sim N(\mu_i, \sigma^2)$$

In keeping with the Generalised Linear Model (GLM) framework, which is well documented, as in (Dobson and Barnett 2008), a single response variable, Y_i is modelled according to two components; a probability distribution for Y_i and an equation linking the expected value of Y_i with a linear combination of explanatory variables. For the situation where the response variable Y_i is not continuous, or where the association between the response variable and the explanatory variable is not in this simple linear form, the model can be generalised for non-linear situations. The model equation can be written in matrix notation:

$$g[E(\mathbf{y})] = \mathbf{X}\boldsymbol{\beta}$$

\mathbf{X} is referred to as the *design matrix* which consists of the known constants denoting the explanatory variables, either categorical levels or measured continuous values. $\boldsymbol{\beta}$ is a vector of parameters representing the change in the response variable for every single unit change in x . The function g is the *link function* which describes the relationship between the mean of the distribution function and the linear predictor. For certain distributions, g can be obtained once the distribution is recognised as belonging to the exponential family.

For a random variable Y , if its probability distribution can be written in the form:

$$f(y; \theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)] \quad (1)$$

Equation 1 Formula for distribution as a member of the exponential family

Where θ is a single parameter, then the distribution belongs to the exponential family. For the distribution to be in the standard form (also referred to as canonical form), a further requirement is that $a(y) = y$. Additionally, $c(\theta)$ must be in a form that causes the probability density function to integrate (or sum for discrete distributions) to one over all values of y . This is referred to as the normalising constant.

In fact, all the probability distributions in the GLM framework belong to the exponential family of distributions (Dobson and Barnett 2008). The binomial distribution caters for binary data where the only two possible outcomes are success or failure. Assuming the probability of success to be given by θ then the binomial distribution probability density function for y of n independent trials is:

$$f(y; \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y} \quad (2)$$

Equation 2 Probability density function for the binomial distribution

This can be written in the exponential family form. The steps for this are given in Appendix A. Based on the form given above, it is easy to see that $a(y)b(\theta) = y \log\left(\frac{\theta}{1-\theta}\right)$ and $c(\theta) = \log(1 - \theta)$ and $d(y) = \log\binom{n}{y}$. Hence, $b(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$ is the link function and this is important for carrying out the appropriate transformation when analysing patient outcome data of various types, as explained in Section 2.1.4.

In summary, given some trial data, the nature of the data will determine the likelihood and the link function $g()$ maps the parameters, γ , that define the likelihood, onto the continuous range $(-\infty, \infty)$.

Chapter 3 Literature Review

This chapter presents a literature review of MAIC methodology and applications, in particular within the context of HTAs. The purpose of this review was to gauge the level of existing understanding in relation to these topics as demonstrated by the published literature. Consideration was given for how current understanding could be strengthened, since at the time of review, it was not obvious that the technique was universally well-understood.

The chapter begins by summarising the basis for using population adjustment methods. This is followed by a general description of the methods before a more detailed look at their technical description. The STC (Simulated Treatment Comparison) method is described here and a brief comparison to MAIC is made. The chapter then describes the review of all available published MAIC methodology including methods and findings. A subsequent review of published applications of MAIC was carried out. These publications are compared in terms of the manner in which the MAIC methodology was described. The results of this review conclude the chapter.

3.1 Basis for Population Adjustment Methods

MAIC and STC are both methods for population-adjusted treatment comparison. As previously described in Chapter 2, they attempt to make the results of two trials more comparable. By reducing the between-trial heterogeneity, the resulting indirect comparisons should be less biased, though there is always potential for unobserved differences to remain (Signorovitch et al. 2012).

The methods rely on the availability of individual patient data, assumed to be drawn from a company's own trial. In a typical commercial context, a pharmaceutical company is interested in performing the MAIC or STC between a treatment that it has developed and a comparator treatment, which may have been developed by a competitor company. Extensive clinical trial data is available for its own treatment including at the individual patient level but such detail on the corresponding clinical trial for the competitor treatment is seldom available to the analyst. Throughout this thesis, the "competitor trial" hereby refers to the one from which the AgD has been drawn (data is not available at the individual patient level). Other scenarios can also arise where a research agency seeks to perform an analysis and has access to some IPD.

The IPD will be characterised by different patient characteristics such as age, or other factors distinctly relevant to the study and treatment area. For instance, this could be whether patients are naïve to a particular treatment or the degree of disease severity. The case of ecological bias in Section 2.1.5 illustrates that use of individual-level data can help to avoid this fallacy. IPD has been used in meta-analysis prior to the introduction of MAIC (Berlin et al. 2002, Donegan et al. 2013, Riley et al. 2010, Stewart and Tierney 2002) as well as in network meta-analysis (Jansen 2012, Veroniki et al. 2016).

The NICE Decision Support Unit have produced a separate document (TSD 3) providing guidance on methods to combat heterogeneity in the relative treatment effect (Dias et al. 2011a). Such heterogeneity is attributed to one of two causes. The first of these is variation in trial settings, protocols or patient populations recruited. The second is due to poor trial quality; if the trial is not double-blinded or where the person allocating patients to treatment arms knows to which arm they will be assigned. Within the document, it is established that network meta-regression using IPD is the “gold standard” approach in terms of attempting to identify true relative treatment effect. MAIC and STC can be regarded as methods that are suited to matching to marginal covariate information (which is generally more commonly available in publications) rather than attempting to match to the full distribution; which would require IPD for the competitor treatment.

3.1.1 Propensity Score Methods

In the field of epidemiology, a common goal is to be able to compare populations or “groups” in terms of incidence of a disease and to be able to generalise for a larger target population. Since sample groups will typically differ in terms of underlying characteristics, raw comparisons of mortality rates are often insufficient. Stratifying mortality rate for numerous subgroups (often age and gender) can overcome population differences, but often the number of subgroups required is too large to manage. In direct standardisation, the mean outcomes for each subgroup are reweighted according to the frequencies noted from the generalised population. However, if few or zero individuals exist for a particular subgroup, it will attract an unnecessarily large weight. Propensity Score methods were applied by (Rosenbaum 1987) resulting in model-based direct standardisation, where, instead of using the observed population frequencies, individuals in each subgroup are weighted according to the inverse of a propensity score. The propensity score is assigned according to the conditional probability that an individual, given their underlying characteristics, is allocated to a subgroup (Rosenbaum and Rubin 1983). The typical setting for propensity scores is concerned with estimating outcomes in a target population based on a sample subpopulation which differs in terms of covariate distribution. Propensity score weighting aims to resolve the differences in the distribution of covariates between the sample and target populations. Individuals in the sample population are attributed a weight based on the inverse of their propensity score.

3.1.2 Calibration

The standardisation methods rely on the sample data being drawn from a subpopulation of the target population. Calibration refers to methods which consider the sample and target populations as distinctly independent (used in two separate clinical trials). In other words, using treatment effect and covariate information from one population to estimate treatment effect in another population characterised by different but known covariate information. IPD is therefore required from both populations. With reference to the situation depicted in Figure 2, MAIC and STC contrast the basis for calibration methods in that the average relative treatment effect between B and C is being

estimated through adjusting the AB population to reflect the target AC population as if treatment B was also administered. In calibration, the relative effect between B and C in the AB population is being estimated (Phillippo et al. 2016). Calibration methods include covariate adjustment, likelihood reweighting, doubly robust methods and the conditional effect estimator.

3.1.3 Introduction to MAIC

The context for which the MAIC method might be used was briefly described in Section 3.1. The MAIC method requires various characteristics to be selected for matching. This process involves the application of individual weights to patients in the IPD dataset such that the summary data of these selected characteristics after weighting will more closely correspond to the AgD available from another trial. In other words, some patients within the IPD trial will now contribute more to the newly calculated means (or other summary statistics) whereas others will contribute less or nothing at all. The anticipated result is that this *matching* of an analyst's own trial data will now make the trial more comparable with the competitor's trial. Any significant relative treatment effect derived from the indirect comparisons should not be attributable to reported differences in the two trial populations.

For a HTA submission, a drug company could provide an alternative estimate of treatment effect using MAIC, even if the relative treatment effect of the new drug subsequently appears diminished compared to that from an unadjusted comparison. In combination with results of an unadjusted ITC, the MAIC will attempt to give an additional perspective on the drug's effectiveness (Proskorovsky et al. 2018). Naturally, the motivation behind this is to reduce bias in the estimate and provide what is hoped to be a fair comparison that more closely reflects the true situation. The basis for this is that the MAIC estimates will correspond to what would have happened if the two treatments being compared were conducted in the same trial with the same patient population. The patient population to which the new estimates are applicable is always that of the competitor trial. However, this only corresponds to the true situation if the patient population is the true target population, which may not be the case and can be difficult to determine.

It should be noted, however, that trial data after population adjustment does not become "gold standard" evidence. Only direct, head-to-head, double-blinded randomised controlled trials will be capable of avoiding the pitfall of confounding (Signorovitch et al. 2011). It is important to also note that variables that are not used for matching may be distorted by the weights applied to the patients. It is therefore beneficial to also report the characteristics that have not been matched, as these can be of relevance (Phillippo et al. 2016). Trial populations may become aligned in some aspects but disjointed in others. At all times, clinical context and relevance should be kept in mind. For example, it may be that certain variables are deemed to have little to no bearing on the particular patient outcome or clinical context may necessitate that certain patient characteristics are confined to a certain range of plausible values.

It is also worth noting a key distinction in population adjustment methods in terms of the anchored and unanchored approach. In an anchored comparison, a common comparator arm exists within the trial data for both the treatment where IPD is available and the target treatment where only summary data is available. From this point onwards unless otherwise specified, treatment B will refer to the treatment for which IPD is available, or the intervention. Treatment C will refer to the case where only AgD is available (such as a competitor's treatment), hereby referred to as the comparator. Finally, treatment A, from here referred to as the anchor treatment, will refer to the treatment that has been directly compared with both treatment B and treatment C in two separate RCTs. These treatments are depicted in the anchored comparison form in Figure 4:

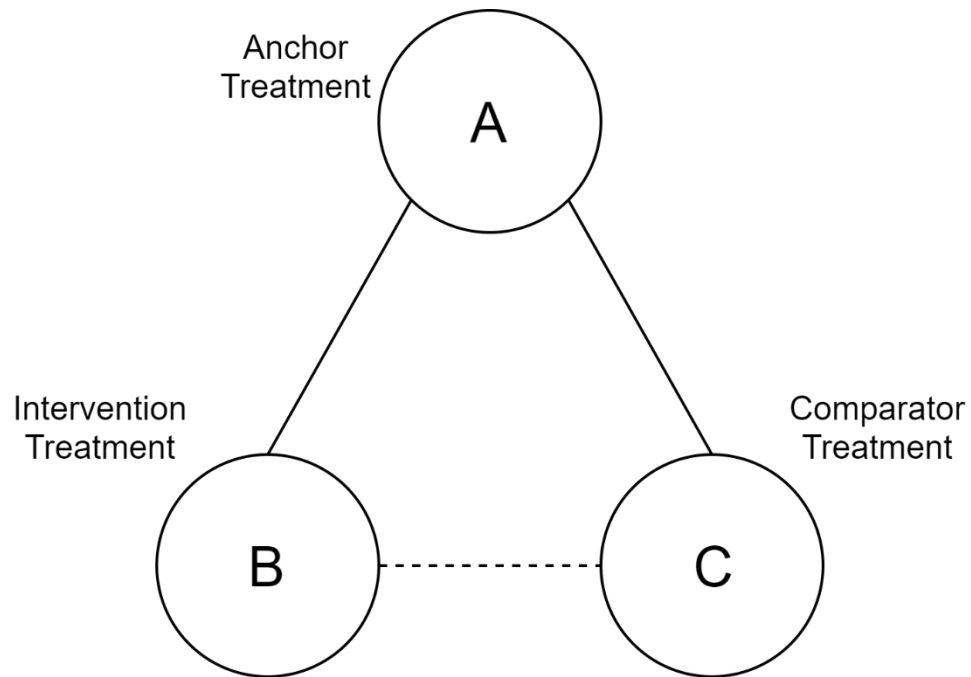


Figure 4 Anchored indirect comparison indicated by dashed line

The indirect comparison between B and C can be said to be ‘anchored’ on common comparator treatment A. The calculation required to make the anchored comparison is depicted as a set of equations in the TSD literature (Phillippo et al. 2017, Phillippo et al. 2018) as follows:

$$d_{BC(AC)} = d_{AC(AC)} - d_{AB(AC)} \quad (3)$$

Equation 3 Anchored indirect comparison

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

Equation 4 Breakdown of anchored indirect comparison

In every subscript, the bracketed component refers to the population and the component outside the bracket refers to the treatment(s) of concern. The trial-level relative treatment effect between the two treatments denoted in the subscripts is represented by d and an estimator of this value is denoted by $\hat{\Delta}$. Hence, $d_{BC(AC)}$ is the relative treatment effect between B and C in the population in trial AC. A trial-level estimator of the expected outcome of treatment C in the AC population or the summary outcome from the trial is represented by $\bar{Y}_{C(AC)}$. The population-adjusted treatment effect of B in the AC population is represented by $\hat{Y}_{B(AC)}$. These values denoting outcome data need to be transformed onto an appropriate scale via function $g()$ as detailed in Sections 2.1.4 and 2.1.5.

Given that in this instance the comparison of treatment effect is made relative to the common comparator for both treatments, the randomisation that is applied in the design of the trials continues to hold for the matching-adjusted indirect comparison (Phillippo et al. 2016). On the other hand, this randomisation property is lost in an unanchored comparison. The unanchored MAIC scenario handles the case where there is a disconnected network of treatments or where single-arm studies are being used. Such trials can be particularly relevant in cases where patient numbers are small, where there are few alternative treatments available, or where a randomising trial would be considered unethical as a treatment benefit is expected. Figure 5 gives a simple illustration of the unanchored comparison below:

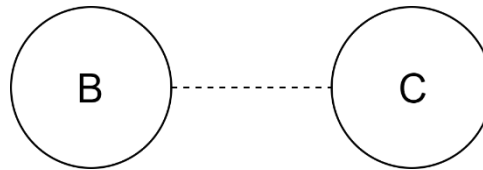


Figure 5 Unanchored indirect comparison indicated by dashed line

Hence, in this situation we are dependent on single-arm trial data. While this is preferable to a naïve unadjusted comparison (Signorovitch et al. 2012), a lack of a comparator arm is a considerable limitation, for which there are implications on the assumptions behind the analysis. This is explored to some length in the Technical Support Document from NICE (Phillippo et al. 2016). Unless stated otherwise, this review will refer to the anchored comparison. The respective equations for the unanchored comparison are as follows:

$$d_{BC(C)} = g(Y_{C(C)}) - g(Y_{B(C)}) \quad (5)$$

Equation 5 Unanchored indirect comparison

$$\hat{\Delta}_{BC(C)} = g(\bar{Y}_{C(AC)}) - g(\hat{Y}_{B(C)}) \quad (6)$$

Equation 6 Breakdown of unanchored indirect comparison

3.2 Simulated Treatment Comparison (STC)

Simulated Treatment Comparison or STC is another method for population adjustment. It involves fitting a regression model of treatment outcome using the IPD from the trial available to the analyst. This results in a model which describes the treatment outcome in terms of a set of covariates and the covariate values available in the IPD should produce an outcome that would be observed if the trial population had a different covariate distribution. As per the situation described in Figure 4, where treatment A is a common comparator, IPD is available for the trial AB but only AgD is available for trial AC, the model for the anchored STC is set out as follows:

$$g\left(\mu_{t(AB)}(\mathbf{X})\right) = \beta_0 + \boldsymbol{\beta}_1^T \mathbf{X} + (\beta_B + \boldsymbol{\beta}_2^T \mathbf{X}^{EM})I(t = B) \quad (7)$$

Equation 7 Formula for simulated treatment comparison

Here, the expected outcome, $\mu_{t(AB)}(\mathbf{X})$ (scaled such that treatment effects are additive via link function $g(\cdot)$) for an individual assigned treatment t with values \mathbf{X} for their patient characteristics (or covariates), is predicted with the following parameters: an intercept term β_0 , a vector of coefficients $\boldsymbol{\beta}_1$ for prognostic variables and $\boldsymbol{\beta}_2$ for the effect modifiers \mathbf{X}^{EM} . The relative effect of treatment B compared to A at $\mathbf{X} = 0$ is given by β_B . The definitions for prognostic variables and effect modifiers were given in Figure 1.

As described in Section 2.1.4, the scale onto which the expected outcome has been transformed is referred to as the linear predictor scale; where treatment effects, effect modification and prognostic variables are additive. This is typically the log odds ratio for binary outcomes, the logit scale for proportions or the log scale for rate outcomes (Phillippo et al. 2016). As noted in Section 2.1.4, the natural outcome scale for binary outcomes is the probability scale. The inverse of the link function, $g^{-1}(\cdot)$, needs to be applied in order to obtain the treatment outcomes on this natural outcome scale.

From Equation 7 above, the population-adjusted estimate of treatment effect for treatment A is given by: $g^{-1}(\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1^T \bar{\mathbf{X}})$ and for treatment B by: $g^{-1}(\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1^T \bar{\mathbf{X}} + \hat{\beta}_B + \hat{\boldsymbol{\beta}}_2^T \bar{\mathbf{X}}^{EM})$. These values are obtained in the R implementation given in Appendix D of the TSD (Phillippo et al. 2016). The way in which the regression model relates to the relative treatment effect for a binary outcome is described in Section 3.5.1.

3.3 Comparison of Population Adjustment Methods

A key difference between MAIC and STC is that whereas the former is better suited to situations where there are few comparators, estimates for many comparators can be obtained with relative ease in STC. This is due to the fact that in MAIC, a set of weights is produced and applied to the available patient data in order to align the treatment's performance to the comparator's population of interest. Therefore, a new set of weights would need to be generated for every comparator and

so MAIC is somewhat more labour-intensive in this instance. In STC, the regression equation can be used to predict outcomes for a comparator treatment when the average values from the comparator trial are substituted into the equation. The same equation in STC can generate the population-adjusted outcome for every comparator once its corresponding average patient characteristic values are substituted in. However, a separate equation will be required for every outcome of interest (Ishak et al. 2015). As a result, it may be preferable to make use of MAIC in the situation where there are multiple outcomes to assess but where there are few comparators. An example of this is illustrated in (Van Sanden et al. 2016). An important point when considering multiple comparators is that the estimates obtained for each are applicable only to a population defined by the comparator summary data. This concludes the comparison between STC and MAIC. It is evident that STC is a versatile and useful method, however, the focus of the remainder of this thesis will be on the MAIC method.

Ishak et al. (2015) describe the key benefits of population-adjusted treatment comparisons. It is stated that these methods can address heterogeneity between studies in a network meta-analysis, can make a single step comparison where a multi-step comparison is otherwise necessary and can also connect an otherwise incomplete network via an unanchored comparison. This approach comes with a number of caveats, which are discussed further in this thesis. The last advantage proposed is that even when a fully connected NMA already exists, it is suggested that new comparisons produced by methods such as MAIC can offer new insights into the effectiveness of a set of treatments. On the other hand, it is worth noting that guidance from the National Institute for Health and Care Excellence does not suggest conducting such comparisons in order to strengthen NMA (Phillippo et al. 2016).

An important aspect of population adjustment to consider is the implied assumption it makes about the representativeness of the comparator trial population. This was a subtle point touched upon in the distinction between methods of calibration and MAIC/STC in section 3.1.2. In Equations 3-6 given previously, the estimates all make reference to a population denoted in the bracketed subscripts. When MAIC/STC is carried out, the population which is adjusted is always the one for which IPD is available; usually the intervention treatment. The target population used for adjustment is assumed to be the one most relevant to the decision question. The estimates derived from MAIC/STC will only be applicable to the target trial population, although the target trial population is unlikely to be the same as the ideal population. The TSD further adds that this situation has previously resulted in an apparent conflict in results between two drug companies; each favouring their own drug in their respective analysis. The real problem is in fact determining which of the trial populations is more relevant for the broader decision question (Phillippo et al. 2016). This further illustrates why it is appropriate for the IPD patient characteristics to be more diverse than those of the AgD set being used to match on. In other words, ranges of patient

characteristics should be wider or the same in the IPD set and sample sizes should ideally be larger.

Overall, the Technical Support Document and more broadly, the general literature, would suggest that the primary goal of the population-adjusted comparison is to account for the heterogeneity resulting from underlying differences in two trial populations. However, these suggestions do highlight the fact that the motivation behind conducting MAIC may fall into one of two main possibilities. Firstly, to make an indirect comparison where it is otherwise impossible to obtain direct evidence between two treatments and attempt to account for population heterogeneity, or secondly, to provide additional evidence and/or an alternative perspective on the true relative treatment effect between two treatments. This could be carried out irrespective of the volume of evidence already available. In this thesis, the first of these is of primary concern, as the literature shows that this is the more common motivation and the implications for HTA are more substantial in this scenario.

3.4 Review Methodology

A core document in the literature review was Technical Support Document 18 (Phillippo et al. 2016), which explored the implications of the MAIC and STC methods and made reference to the key article proposing the MAIC method (Signorovitch et al. 2010). The report concludes with key recommendations on the usage of the methods, as guidance for HTA in NICE. The same TSD group produced an overview of population-adjusted treatment comparison methods concisely laid out on a poster, which was presented at ISPOR's 22nd Annual International Meeting (Phillippo et al. 2017).

Subsequent to the original article proposing MAIC, Signorovitch et al. (2012) wrote an article summarising the method and exploring the contributions that the method makes to comparative effectiveness research, using example applications for illustration. A number of existing MAIC applications and broader evidence synthesis methodology papers were then considered. The remaining literature examined in this review was identified through references from the initial set of papers. Where additional MAIC examples were sought, basic searches were run using keywords such as “matching adjusted indirect comparison” on the PubMed platform and Embase database. In addition, alerts were created to inform of any new publications appearing on these sites that were published during the course of the project. These databases cover a wide variety of publications such as Value in Health. This search used variants on this basic terminology as well as “simulated treatment comparison”. Observations showed that during the period of the review, a new publication making use of this terminology would be uploaded to PubMed at an average rate of approximately one every two weeks. However, many of the observed uploads were updates to existing applications as opposed to novel analyses. It is important to also acknowledge that such keyword-based searches are always subject to limitations and may not capture the full scope of

the material available. This is especially relevant to MAIC and STC, due to the broad range of terminology used, particularly in earlier publications. This is described in Section 3.6.3.

To complement the literature search, a search for existing R packages uploaded to CRAN was performed using (METACRAN 2019). No R packages that could perform a population adjustment method were discovered, however there were a small number of existing packages designed to support network meta-analysis. A brief note on these is included at the end of the review.

3.5 MAIC Methodology

Signorovitch et al. (2012) give the rationale for the MAIC method at a non-technical level. In particular, using example applications to show how use of IPD can resolve the various limitations that may arise from indirect comparisons. In the article, the MAIC approach is summarised in three steps:

1. Conducting a systematic review, where clinical trials are identified to compare the relevant treatments.
2. Identifying a precise definition of an outcome measure which can be compared across all included trials and therefore the different treatments being considered. Variation can be present in multiple aspects of the outcome measure definition.
3. The weighting aspect of the matching-adjusted indirect comparison is the third key part of the method, the detail of which is covered in (Signorovitch et al. 2010) and is outlined in this review. As an additional step, patients who could not be enrolled in the trial for which only aggregate data is available (the competitor's), should also be excluded from the IPD that is available. For example, patients might be excluded for being naïve to a particular treatment. Selecting data in this way could have a significant impact on the analysis as we are effectively discarding data, but the idea is to rule out patients that would have been definitively ruled out of the competitor's trial due to the recruitment criteria (also referred to as inclusion/exclusion criteria).

Patient characteristics must first be selected for matching. In an anchored MAIC, this means accounting for all effect modifiers. As noted in Section 2.1, when we examine the covariates of a given dataset, a purely prognostic variable is one which does not affect the relative treatment effect, even for different levels of the covariate. That is to say that while absolute outcomes may be affected when we vary the covariate level, the improvement or deterioration should be the same across the treatments being compared. In an anchored scenario, since there is data available for the same anchor treatment in the two trials, although the average level of the covariate may be different between them, the anchor treatment acts as a common reference arm for both treatments of interest. However, an effect modifier refers to the situation where there is an interaction between a covariate and a treatment outcome. Since the treatment effect may vary for particular subgroups

for one or some of the treatments being compared, this has an impact on relative treatment effect. Hence, effect modifiers must be selected for matching.

In the unanchored comparison, the evidence is no longer randomised as we are using distinct single-arm trial data. As a result, there is no guarantee that the level of each prognostic variable remains the same among single-arm studies. There is no randomisation taking place in order to balance subgroups of each covariate between the two treatments. A treatment could appear more effective as a result of a more favourable patient population in the trial who are more receptive to treatment. This means that we cannot avoid the possibility of some form of bias influencing the analysis. Therefore, prognostic variables must also be matched on for the unanchored case. In addition to this, the NICE TSD states that the analysis should be clearly pre-specified. Selecting a variable to match, based on an analysis that was already conducted revealing that the variable was an effect modifier is referred to as “post hoc reasoning” and is to be avoided (Phillippo et al. 2016). There are three main methods for identifying effect modifiers and prognostic variables: consultation with clinical experts, review of epidemiological literature and MAIC publications in search of relevant literature, and regression analyses using the available IPD (Kamangar 2012, National Institute for Health and Care Excellence 2013, Phillippo et al. 2016).

To explain the basis of the MAIC methodology, an example scenario will be defined. Suppose there is IPD available for a treatment $t = 0$ (our own trial), but only AgD is available for $t = 1$ (competitor’s trial). We want to reweight the IPD from $t = 0$ to match the AgD characteristics for $t = 1$. Assume there are only two covariates: age is an effect modifier and gender is a prognostic variable. For an individual patient i who is assigned to treatment t and has baseline characteristics \mathbf{X}_{it} , the propensity score weight which will be applied to the patient is the inverse odds of them being in the trial with IPD available versus the other trial. This is represented by:

$$w_{it} = \frac{Pr(t_i = 1 | \mathbf{X}_{it})}{Pr(t_i = 0 | \mathbf{X}_{it})}$$

The probabilities indicated here refer to the probability of the patient enrolling onto treatment $t = 1$ or $t = 0$ respectively i.e. their ‘propensity’ for one or the other, in this case, based on their age. Any patients that were more likely, given their age, to enrol into trial $t = 1$, will be upweighted. Conversely, patients that were less likely to enrol into trial $t = 1$ due to their age, will be downweighted as they are over-represented prior to the application of the weights. The result of this is that the means of the baseline characteristics of our IPD after applying weights will more closely resemble those of the AgD. Other summary statistics can also be matched. Falling short of being able to match to the full covariate distribution (for which other methods are more appropriate subject to the provision of IPD on all trials, see Section 3.1), it may be possible to match to a median, a range, various quartiles, correlations or higher moments. In practice, these types of

distributional information are not used for matching because often, “only the marginal mean and standard deviation of each covariate is known” (Phillippo et al. 2016). In developing a package in the R programming language to implement the method, matching to a median statistic proved to be a considerable challenge, which may not have been adequately addressed in the literature. This issue is fully described in Section 4.4.2.

Logistic regression is typically used to develop propensity scores (Olmos and Govindasamy 2015). Thus, for an anchored comparison, the weights are calculated through estimating a logistic propensity score model that includes all effect modifiers but no purely prognostic variables:

$$\begin{aligned} \log(w_{it}) &= \alpha + \beta X_{it} \\ w_{it} &= e^{\alpha + \beta X_{it}} = e^{\alpha} e^{\beta X_{it}} \end{aligned}$$

Maximum Likelihood Estimation (MLE) would normally be used to estimate β or other parameters of this model. However, that is not possible here because there is no IPD for $t = 1$. In order to find β , we require weights that are estimated such that they lead to matching of the effect modifier distributions between the trials for treatments $t = 0$ and $t = 1$. In other words, for the anchored scenario, we desire the situation where the mean of the weighted IPD is equal to the mean of the baseline characteristics for the AgD. The mean of the weighted IPD can be represented as follows:

$$\frac{\sum_{i:t=0} x_{it} \hat{w}_{it}}{\sum_{i:t=0} \hat{w}_{it}} = \frac{\sum_{i:t=0} x_{it} e^{\alpha} e^{\hat{\beta} X_{it}}}{\sum_{i:t=0} e^{\alpha} e^{\hat{\beta} X_{it}}} = \frac{\sum_{i:t=0} x_{it} e^{\hat{\beta} X_{it}}}{\sum_{i:t=0} e^{\hat{\beta} X_{it}}} \quad (8)$$

Equation 8 Mean of the weighted IPD

For the unanchored scenario, the t index can be omitted from Equation 8 as we will no longer be comparing two-arm trial data involving a placebo arm for each of the two treatments. The mean of the weighted single-arm IPD will be matched to the mean given from the single-arm trial of the competitor treatment. The mean of the weighted IPD is set to equal $\bar{X}_{(t=1)}$:

$$\begin{aligned} \frac{\sum_{i:t=0} x_{it} e^{\hat{\beta} X_{it}}}{\sum_{i:t=0} e^{\hat{\beta} X_{it}}} &= \bar{X}_{(t=1)} \\ \frac{\sum_{i:t=0} x_{it} e^{\hat{\beta} X_{it}}}{\sum_{i:t=0} e^{\hat{\beta} X_{it}}} - \bar{X}_{(t=1)} &= 0 \end{aligned} \quad (9)$$

Equation 9 Equalising the weighted IPD with the AgD for comparator treatment

To find a unique, finite solution for $\hat{\beta}$, note that:

$$\sum_{i:t=0} x_{it} e^{\hat{\beta} x_{it}} - \bar{X}_{(t=1)} \left(\sum_{i:t=0} e^{\hat{\beta} x_{it}} \right) = 0$$

The above equation can be re-written in terms of $\sum_{i:t=0} e^{\hat{\beta} x_{it}}$

$$\left(\left[\sum_{i:t=0} x_{it} \right] - \bar{X}_{(t=1)} \right) \sum_{i:t=0} e^{\hat{\beta} x_{it}} = 0$$

$$\sum_{i:t=0} x_{it} e^{\hat{\beta} x_{it}} - \sum_{i:t=0} \bar{X}_{(t=1)} e^{\hat{\beta} x_{it}} = 0$$

$$\sum_{i:t=0} (x_{it} - \bar{X}_{(t=1)}) e^{\hat{\beta} x_{it}} = 0$$

Without loss of generality, it can be assumed that $\bar{X}_{(t=1)} = 0$, implying that:

$$\sum_{i:t=0} x_{it} e^{\beta x_{it}} = 0 \tag{10}$$

Equation 10 Derivative for method of moments to estimate β (assuming that $\bar{X}_{(t=1)} = 0$)

Equation 9 is therefore equivalent to Equation 10. Setting Equation 10 as a derivative; the objective function is:

$$Q(\beta) = \sum_{i:t=0} e^{\beta x_{it}}$$

$Q(\beta)$ is convex (verifiable from $Q''(\beta)$) and so the minimum of this function is a unique solution, our method of moments estimate, $\hat{\beta}$.

We therefore want to minimise this objective function while retaining the condition that $\bar{X}_{(t=1)} = 0$. For this to be satisfied, the matrix of effect modifiers, \mathbf{X}_{it}^{EM} is centred on $\bar{X}_{(t=1)}$ i.e. the subtraction $\mathbf{X}_{it}^{EM} - \bar{X}_{(t=1)}$ is applied. This centring can be done without loss of generality and is straightforward to perform for covariates where the target statistic to match to for a covariate is a mean. On the other hand, where matching to a standard deviation is desired, since this is a function of variance, the second moment of the covariate also needs to be centred. For the second moment, note that: $E(X^2) = E(X)^2 + Var(X)$, hence we subtract the sum of the target mean squared and standard deviation squared from the second moment of the covariate in the IPD.

The notation in the literature may refer to $\boldsymbol{\beta}^T$, a vector, with one element for each effect modifier which is transformed such that the matrix multiplication $\mathbf{X}_{it}^{EM} \boldsymbol{\beta}^T$ is possible. The columns in \mathbf{X}_{it}^{EM} will therefore equal the number of rows in $\boldsymbol{\beta}^T$ and this dimension will correspond to the number of effect

modifiers being weighted (as well as any second moments required). Simpler notation has been used for these equations in order to illustrate that for every effect modifier, a single value for β is required.

Having assured that $\bar{X}_{(t=1)} = 0$ by centring the relevant covariates in the IPD, the objective function can then be minimised using a minimisation algorithm. The BFGS algorithm is used in the Technical Support Document (Broyden 1970, Fletcher 1970, Goldfarb 1970, Shanno 1970).

This will produce an estimate of β via the method of moments. In essence, we are finding a value for $\hat{\beta}$ such that for patients where $t = 0$, they will be re-weighted by $\hat{w}_{it} = e^{\hat{\beta}X_{it}}$. With the value for $\hat{\beta}$, the weights (\hat{w}_i for each i th patient) to be applied to the IPD can now be estimated. Together, the summary statistics of the patient characteristics for the newly weighted patients will match the baseline characteristics in the AgD for treatment $t = 1$.

Note from the distribution of the weights $\hat{w}_{it} = e^{\hat{\beta}X_{it}}$, that they will be greater than but not equal to zero, though may be infinitesimally close to zero. When the weight drops to such a low value, we can consider the corresponding patient to be, in effect, excluded from the analysis. This leads to an effective sample size (ESS) for the data after weighting, which represents the number of patients actually included in the analysis after weighting. If this is small, then there is less overlap between the two trial populations and the indirect comparison is inherently less valid. In general, including more baseline characteristics as matching parameters leads to a small ESS. An approximation for the effective sample size can be calculated from the following:

$$ESS \approx \frac{\sum_{i,t} (\hat{w}_{it})^2}{\sum_{i,t} \hat{w}_{i,t}^2}$$

Finally, in order to simplify presentation, if we ignore the placebo arm data, the matching-adjusted treatment effect estimate can be obtained by substituting these weights into the equation:

$$\begin{aligned} \hat{\theta} &= \frac{\sum_{i=1}^n y_i (1 - t_i) \hat{w}_i}{\sum_{i=1}^n (1 - t_i) \hat{w}_i} - \bar{y}_1 \\ &= \frac{\sum_{i:t_i=0} y_i e^{\hat{\beta}X_{it}}}{\sum_{i:t_i=0} e^{\hat{\beta}X_{it}}} - \bar{y}_1 \end{aligned}$$

Here \bar{y}_1 refers to the mean outcome for $t = 1$. For an anchored comparison, let $z = 0$ represent the placebo arm and $z = 1$ represent the treatment arm, the aggregate baseline and outcome data for the placebo arm is represented by $\bar{x}_1^{(0)}$ and $\bar{y}_1^{(0)}$ respectively, and $\bar{x}_1^{(1)}$ and $\bar{y}_1^{(1)}$ for the treatment arm (Signorovitch et al. 2010).

The estimated relative treatment effect from the MAIC is then given by:

$$\hat{\theta} = \frac{\sum_{i:t_i=0,z_i=1} y_i e^{\hat{\beta}x_{it}}}{\sum_{i:t_i=0,z_i=1} e^{\hat{\beta}x_{it}}} - \frac{\sum_{i:t_i=0,z_i=0} y_i e^{\hat{\beta}x_{it}}}{\sum_{i:t_i=0,z_i=0} e^{\hat{\beta}x_{it}}} - (\bar{y}_1^{(1)} - \bar{y}_1^{(0)})$$

This equation is of the same form as that from the TSD literature (Equation 3 in this thesis). In practice, this estimated relative effect between two treatments would be calculated rather than estimates of absolute outcomes. Estimates of absolute outcomes will be biased unless if all prognostic variables and effect modifiers that are not balanced across the populations can be accounted for. It is also easier to calculate the population-adjusted treatment effect using a simple linear model instead of using the weighted means and this gives an equivalent result. The linear model is explained in the next section.

3.5.1 Relating the Logistic Model with Relative Treatment Effect

In estimating relative treatment effect between two treatments recording a binary outcome, our aim is to predict or generalise the occurrence of the event based on the treatment arm to which the patient was assigned. As stated previously, relative treatment effect is defined as the difference in the average treatment effects between two treatments, both subject to scale conversion where required.

In the example implementation in the Technical Support Document appendix, given in R, it is indicated that the relative treatment effect for B vs. A adjusted for the population in trial AC can be estimated using a simple linear model (Phillippo et al. 2016). In the case of a binary outcome this makes use of the `glm()` function where the family argument is set to `binomial`. The advantage stated in the document of using the linear model as opposed to calculating the odds ratio by hand is that it allows for standard errors to be calculated using a “sandwich estimator” (White 1980). The logistic model produced will return an intercept value and a coefficient value for a reference group i.e.

$$y = \text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

$$y = \beta_0 + \beta_1 I(\text{treatment}B) \quad (11)$$

Equation 11 Logistic model equation for binary patient outcome

When applying the GLM, the outcome, y (in the binary case, whether the patient experienced the event or not), is being predicted based on whether the patient was in the treatment arm or the common comparator arm. This is represented by the $I(\text{treatment}B)$ indicator variable term in Equation 11 (in this case, taking value 1 for the B intervention and 0 for the common comparator, A). The event outcome is therefore regressed on the treatment group.

As a result, for a patient in the common comparator A arm, the model returns the intercept term as the outcome. The intercept term is therefore equivalent to the log odds of treatment A or $g(\bar{Y}_{A(AB)}) = \ln\left(\frac{\sum_{i=1}^{N_A} Y_A}{N_A - \sum_{i=1}^{N_A} Y_A}\right)$. Here the term $\sum_{i=1}^{N_t} Y_t$ refers to the sum of the outcome values for N_t patients that were administered treatment t . The summary outcome obtained from the trial for treatment t in the AB population is represented by: $\bar{Y}_{t(AB)}$. The model returns the outcome for the intervention treatment B by adding the difference between the two treatments' log odds, in other words, the log odds ratio between treatment B and treatment A or the relative treatment effect obtained from the trial:

$$\hat{\Delta}_{AB(AB)} = g(\bar{Y}_{B(AB)}) - g(\bar{Y}_{A(AB)}) = \ln\left(\frac{\sum_{i=1}^{N_B} Y_B}{N_B - \sum_{i=1}^{N_B} Y_B}\right) - \ln\left(\frac{\sum_{i=1}^{N_A} Y_A}{N_A - \sum_{i=1}^{N_A} Y_A}\right)$$

Standard errors can be calculated using a sandwich estimator, which, in the TSD is noted to be typical for MAIC, but bootstrapping (as used in (Warren et al. 2018)) and Bayesian methods may also be used. It is worth highlighting that different approaches to calculating standard error may have some implication for MAIC analyses in HTA. Introducing variation in the derivation of uncertainty for MAIC estimates presents a challenge for fair comparison of treatments in HTA. This is compounded by the fact that HTA models are typically nonlinear; Markov models result in outputs that are a result of multiplicative and power operations (Briggs et al. 2006). When passing parameter values through into a model, the expected value of the outputs of interests (for instance, the total costs of an intervention) will not be obtained simply by passing in the expected value of the various input parameters i.e. the point estimates obtained from MAIC or NMA. Instead, a typical sensitivity analysis will involve specifying a range of values for the input parameters according to an appropriate distribution. This leads to a distribution of output parameters. While these ranges may not themselves be so broad, introducing different methodologies for estimating uncertainty could lead to a substantial range of available estimates. Besides estimating uncertainty, other variations in the MAIC technique are described in the ensuing section.

3.5.2 Variants of MAIC Methodology

It is noted in the TSD that the MAIC method, as proposed in the original methodology articles (Signorovitch et al. 2012, Signorovitch et al. 2010), estimates weights for the entire AB population instantaneously (Phillippo et al. 2016). Alternative weighting techniques are present in the literature (Petto et al. 2018) as well as entirely different methodologies proposed for indirect comparison such as polynomial weighting (Regnier et al. 2016).

It is also worth highlighting the variation in the literature surrounding the technical details for implementing the method. The implementation from the TSD demonstrates how to apply MAIC using exact matching in R via BFGS minimisation (Phillippo et al. 2016) whilst the appendix to the

methodology article refers to Newton-Raphson optimisation (Signorovitch et al. 2010) implemented in SAS. Although no code is provided, a separate paper presents an implementation of the method in SAS, but through adopting a sampling approach (Malangone and Sherman 2011). The proposal involves drawing 1000 subsets of the IPD with each having a distribution corresponding to the required AgD. This is done using the SURVEYSELECT procedure in SAS. The paper then goes on to estimate median survival for each of the 1000 samples using LIFETEST. Following this, an ODS statement is used to combine these survival estimates together. This results in bootstrapped sampling where 95% confidence intervals can be calculated.

Malangone and Sherman (2011) have looked at the case of time-to-event data with the aim of estimating median survival. In analysing the effectiveness of treatments in oncology and certain other therapeutic areas, it is more typical to analyse time-to-event outcomes. This form of analysis changes the focus from probability of experiencing a certain event to the length of time before experiencing the event. Where the event of interest is death, time-to-event data refers to survival time. More specifically, the length of time may relate to the first point of observed relapse or progression of the disease and hence more specific endpoints may be used such as progression-free survival as opposed to overall survival.

Survival time outcomes present an added complication in that it is not as straightforward to publish the necessary summary statistics as with continuous or binary outcome data. Survival data often includes censoring, where information on the survival of the patient is terminated prior to the patient experiencing the event of interest. Therefore, a calculation of average survival time becomes biased. Such data is therefore analysed in the framework of survival analysis; a separate branch of statistics. It is recommended that hazard ratios be used as the effect measure for clinical studies (Higgins and Green 2011). With the provision of Kaplan Meier curves, IPD can be reconstructed using an algorithm from (Guyot et al. 2012).

The existence of the methodological variations described above has implications for the results that ensue from any MAIC. In a poster presented at ISPOR's 20th Annual Conference, (Wang et al. 2015) explored Malangone and Sherman's method of resampling bootstrapping to perform MAIC. Their findings showed that in the presence of an interaction between baseline characteristics and the treatment, the resampling technique returned opposing conclusions regarding which treatment was more effective, depending on the treatment for which IPD was available. Wang et al. (2015) suggest further research to identify when the resampling approach may be best applied. The example of time-to-event data in (Malangone and Sherman 2011) has also highlighted that analysis of such data entails a number of extra steps before the adjustment can take place. The scope of the problem of variation in methodology is summarised in Section 3.6.3, in light of published descriptions of the method obtained from a review of published MAICs.

3.5.3 Key Assumptions Underlying the Methods

In a standard indirect comparison or fixed effect NMA, it is assumed that the relative effect A vs C in an AC trial is identical to the A vs C effect that would be expected in the trial AB (consistency). This is referred to as *constancy of relative effects* on the linear predictor scale; which refers to the scale on which treatment effects are additive. For this to hold, any differences in the relative effects need to be accounted for entirely by an imbalance in the effect modifier variables (these may be denoted by X^{EM}) (Phillippo et al. 2016).

For an anchored population-adjusted indirect comparison, this assumption is relaxed somewhat such that the relative effect observed from the AC trial at a given covariate level (for example, Body Mass Index = 23.0) is equal to that from the AB trial. This means that only effect modifiers need be adjusted for, whereas prognostic variables are already balanced between studies. This is referred to as *conditional constancy of relative effects* on the natural outcome scale (Phillippo et al. 2016); the scale on which the outcome is defined and observed. Effect modifiers may already be balanced between trials but may become unbalanced after reweighting in MAIC. Therefore, the anchored MAIC needs to include effect modifiers that are both in balance and imbalance between trials. However, unlike in STC, there is little value in including prognostic variables in the weighting model since this will reduce the effective sample size.

The Technical Support Document indicates that for STC, it is important to specify a correct outcome model in order to obtain unbiased estimates. Therefore, it is recommended that, in addition to effect modifiers in imbalance, that any other prognostic variables or effect modifiers (that begin in balance) should be included if they improve model fit.

For the unanchored population-adjusted indirect comparison, *conditional constancy of absolute effects* is assumed. This means that the differences between the absolute outcomes that would be observed in each trial are caused by imbalances in both prognostic variables and effect modifiers. It is highly unlikely that all the prognostic variables and effect modifiers would be known; this is a difficult assumption to meet. Knowing all these imbalances would, in effect, make conducting randomised controlled trials redundant. Furthermore, since unanchored comparisons have this additional requirement of all prognostic variables in imbalance being accounted for in the weighting for MAIC, these comparisons will always be less precise than the anchored MAIC (Phillippo et al. 2016). Similarly, in STC, all effect modifiers and prognostic variables must be included in the outcome model and thus the requirement to correctly specify the outcome model becomes more onerous to achieve.

Correctly specifying an outcome model in STC becomes especially important when it is defined on a transformed linear predictor scale; where treatment effects, effect modification and prognostic variables are additive, but the indirect comparison is performed on the unchanged natural outcome scale. An indirect comparison made through STC that makes use of a link function to transform

data between the two scales assumes that all effect modifiers and prognostic variables are accounted for and importantly, with respect to the linear predictor scale. By contrast, MAIC has the advantage of not requiring a fixed outcome scale to be defined for the weighting model (although the indirect comparison does assume additivity on a specific scale).

3.5.4 Recommendations from Technical Support Document 18

The authors of the TSD conclude that population-adjusted indirect comparisons have a role to play in HTA but, particularly with such a limited amount of research available on the methods, their use should be justified on a case-by-case basis (Phillippo et al. 2016). Throughout this review, reference has been made to TSD 18 as a key article of reference in respect of the methodology. The TSD's authors make a number of other contributions which have been discussed in this thesis. These include the issue of scale in carrying out indirect comparisons, the importance of careful selection of covariates to match on rather than including as many as possible and the importance of the target population in using these methods. The document sets out a clear list of recommendations on the use of these population adjustment methods as well as guidance on their use in HTA submissions to NICE. They can be summarised as follows:

1. When a connected network is not possible, only then consider an unanchored indirect comparison (single-arm studies).
2. It must be shown that there are effect modifiers present and that population adjustment would remove substantial bias.
3. In the case of an unconnected network, unanchored comparisons must include assessments of error as a result of covariates unaccounted for.
4. All effect modifiers to be matched, even if they are already balanced.
5. Indirect comparisons to be performed on the *linear predictor scale*.
6. The target population must be stated.

This list is followed by a set of reporting recommendations for an MAIC/STC analysis. With consideration for the points above, the TSD states that HTA submissions should, in addition to reporting of results, include measures of uncertainty. The distributions of the covariates before and after matching should also be presented as well as evidence for effect modifier or prognostic variable status. In the case of MAIC, the distribution of weights applied should also be presented with a histogram.

At the end of their methodological summary, the TSD group make the following conclusion in respect of the use of MAIC and STC in technology appraisals: "in the interests of transparency and consistency, and to ensure equity for patients and a degree of certainty for those making

submissions, it is essential to regularise how and under what circumstances these procedures should be used” (Phillippo et al. 2016). Moreover, in a subsequent, recent publication reviewing technology appraisals submitted to NICE that involved population adjustment methods, further research was recommended in the form of simulation studies (Phillippo et al. 2019). This requirement is supported further by the following review of published applications of the MAIC method.

3.6 Review of MAIC Applications

In order to more explicitly gauge the level of existing understanding of the MAIC method that was evident in the literature, a review of MAIC applications was conducted. Of interest here was the way in which the method was described in each of the studies and consideration for whether a reader with little to no knowledge of the method could be expected to understand the steps that were being taken in the analysis. In terms of the selection process, it was preferable to capture the full range of diversity across applications in terms of: journal of publication, year of publication, company conducting the analysis, therapeutic area and to also capture different method variations. Details of the full range of publications included for review are listed in Appendix B. It was with this diversity in mind that certain publications were selected for focus. Five of these published articles are covered in this review in greater detail. These are:

- Van Sanden et al. (2016): simeprevir + peginterferon alfa 2a + ribavirin (IPD) vs peginterferon alfa 2a + ribavirin (AgD)
- Regnier et al. (2016): ranibizumab (IPD) vs aflibercept (AgD)
- Van Sanden et al. (2018): daratumumab (IPD) vs pomalidomide + dexamethasone (AgD)
- Odom et al. (2017): sonidegib (IPD) vs vismodegib (AgD)
- Warren et al. (2018): ixekizumab (IPD) vs secukinumab (AgD)

Other articles that were reviewed are explored briefly here, before the main findings from the articles listed above are detailed.

There were a number of challenges in performing the review of MAIC applications. A significant challenge is that the terminology used in these studies may not have included “MAIC” or “Matching-Adjusted Indirect Comparisons” at all, as was the case for (Regnier et al. 2016). Other studies (which were not detailed in this review) may have indicated relevant keyword terminology such as “comparative effectiveness” and “matching” but would instead be making use of a Simulated Treatment Comparison, such as in (Sivaprasad et al. 2016). In one early example paper, the approach was elaborately described as a “propensity score matching to assemble cohorts in which...” “...patients were balanced on baseline characteristics” and the probability of the patient enrolling onto the trial of interest “was estimated for each patient using a nonparsimonious multivariate logistic regression model” (Carlin et al. 2013). The varied manner in which the method

was described across publications warranted thorough reading of each study to ensure that the method conformed to the established protocol for MAIC or that method variations could be suggests that full capture of MAIC applications in the literature may be challenging.

3.6.1 Overview of the Literature

Beyond the method application presented in (Signorovitch et al. 2010), while there were a number of published studies making use of the MAIC method, few of these gave detailed outlines of the methodology employed. This is evident in (Atkins et al. 2019) where greater emphasis is made on the results of the MAIC analysis. This may have been due to the type of journal in which the study was published; in this case, a specialist medical journal in immunotherapy, as part of the Future Medicine journal portfolio. It is primarily concerned with scientific advances in the field (Immunotherapy 2019). Irrespective of this, it is difficult to verify or critically appraise this application of MAIC. This is a recurrent theme in the literature. In the case of (Song et al. 2019), the authors relied heavily on referencing the methodology paper. Although references were usually made to the methodology papers of Signorovitch et al. (2010, 2012), in general, most articles would simply refer to the context of lack of comparative data, the use of indirect comparisons, the need to account for differences in population characteristics and a basic description of MAIC. Therefore, unless referenced otherwise, it seems likely that the method of moments approach as described in (Signorovitch et al. 2010) was being employed. It would be desirable for all published studies to make this choice explicit, since, among the studies, there was a considerable lack of detail in this regard. An outline of the method used and in particular drawing attention to any variations of this would be useful in every application of MAIC.

Some authors from within the literature focus on the usefulness of the unanchored approach (Majer et al. 2017, Van Sanden et al. 2016), though sometimes without due acknowledgement of its limitations as were described in the TSD (Phillippo et al. 2016). The application presented by (Signorovitch et al. 2015) is better described. Here, the authors note that indirect comparison methodology has evolved such that their study offers a superior analysis to a previous study comparing the same two treatments (nilotinib and dasatinib) which did not use a common comparator. Signorovitch et al. (2015) also comment on the limitations of MAIC analysis. One critique of this publication is that no mention of ESS is made in the article and in the table of baseline characteristics, while the adjustments appear relatively small owing to the similarity between the trials, the same figure for n is reported for before and after matching which is concerning.

An update to an older, pre-existing MAIC analysis is presented by (Ishak et al. 2018) and is introduced with a thorough background to the method. Here the perspective is switched since the treatment for which IPD was available (sunitinib) had been the opposing treatment in the original analysis carried out by (Signorovitch et al. 2013). Ishak et al. (2018) include a comparison between

the Bucher method, the anchored as well as unanchored matching-adjusted indirect comparison. This is nicely displayed in a table with the comparison contextualised for the application to hazard ratios quantifying progression-free survival as well as overall survival (Ishak et al. 2018). By contrast, the first analysis by Signorovitch et al. (2013) gives little background description to the method. A key divergence from other publications is that (Signorovitch et al. 2013) refer exclusively to matching to medians as opposed to means. The median of age was matched while all the other characteristics matched on were proportions (Signorovitch et al. 2013). There is additionally a characteristic measuring the proportion of patients above the age of 64. It is not explained why this age or subgroup was selected for matching. It is possible that matching solely to the median of age was insufficient, since matching to a median is not as straightforward for reasons detailed in Section 4.4.2 which concerns implementation of matching using R.

In their article, (Signorovitch et al. 2013) draw attention to the benefits of using MAIC to address common challenges arising in comparative effectiveness research in new oncology treatments. These are identified as: the small numbers of trials from which to draw data and the bias arising from patient crossover from placebo arm to an active therapy arm. It is common practice to reassign a patient to a different treatment in the course of the trial period. This is referred to as a crossover and entails that the placebo trial arm is “contaminated” by crossovers to an active therapy following disease progression. As a result and as performed in this analysis, there is a trend for an anchored MAIC to be carried out for a progression-free survival outcome and an unanchored analysis to be used for the overall survival outcome. Signorovitch et al. (2013) comment on the general limitations of their analysis, particularly given this challenge of limited trial data involving crossover patients, however, they do not make any reference to effective sample size and this is a key component to any MAIC analysis.

Tremblay et al. (2018) describe the basis for MAIC well in terms of indirect treatment comparisons. They state that MAIC “has provided strong comparative evidence in the absence of head-to-head studies in various disease settings” (Tremblay et al. 2018). The article does not support this assertion other than through citation of past MAIC publications, however, it is reasonable to suggest that MAIC has served to offer a new perspective on the effectiveness of various treatments. The technical description given for MAIC is somewhat less effective. The authors state that the weights are “created by performing a logistic regression on the patient-level RIBO data that included an extra observation representing the comparator’s data...” and “The predicted values (or propensity score) that resulted from the logistic regression were used to weight the RIBO trial data”. While it is stated that the output values are used to weight the trial data, this description could be confused with the STC method, which uses logistic regression to predict treatment outcomes directly. In MAIC, the focus is on the weighting of baseline characteristics and outcomes to perform the adjustment, not on the logistic regression used to estimate the weights. One positive

attribute to this paper is that the statistical software and TSD guidelines are cited. On the other hand, the study included an unanchored as well as anchored analysis with both sets of outcome results reported and it is not clear why the unanchored analysis was performed.

Proskorovsky et al. (2018) explain the concept of ESS including the danger of low ESS. Despite one ESS in the initial base-case analysis being as low as approximately 30% of the original sample size (61 from 194 patients) (Proskorovsky et al. 2018), no acknowledgement of this is made in the limitations section or elsewhere. As a matter of presentation, two ESS figures were given for each post-matching column, separated by a slash. The reader would need to carefully note that these are the ESS figures relating to the base case and sensitivity analyses and do not represent a fraction or a comparison against original sample size.

3.6.2 Findings from Selected MAIC Applications

Van Sanden et al. (2016) present a special application of MAIC for examining the effectiveness of combining a therapy to an existing treatment for a patient afflicted with genotype 4 hepatitis C viral infection. The standard therapy was peginterferon alpha 2a + ribavirin (PR) and was evaluated against combining simeprevir with this therapy. In other words, simeprevir was not being proposed as a standalone therapy to substitute the existing treatment available. This is an example of an unanchored MAIC where “it was necessary to estimate the incremental benefit that simeprevir offers in genotype 4 infected patients, over and above that achievable with PR alone” (Van Sanden et al. 2016). IPD from the newer single-arm study examining simeprevir in addition to peginterferon alpha 2a + ribavirin (RESTORE trial) was used to match against various separate single-arm studies examining PR alone.

Van Sanden et al. (2016) state that MAIC is a particularly useful method thanks to its unanchored approach, making treatment comparisons possible without a common comparator arm. They describe the process of systematic review for studies of relevance and what they implicitly refer to as the first stage of matching in terms of excluding patients from RESTORE that would have been excluded from the trials that studied PR alone. After the “second stage” of matching where the baseline characteristics are matched, there is a discussion of which of the five studies are most comparable. It is shown that only two were able to match on a sufficient number of matching parameters; the remaining three could only match on two parameters (fibrosis level and baseline viral load). Nonetheless, five separate unanchored MAICs were performed. The authors state that they “assigned an additional requirement that at least two parameters of high relevance (in this case fibrosis stage and viral load) should also be matched”. While these covariates were identified as being of high relevance to treatment response by two hepatologists, it would be preferable to confirm why this particular requirement was assigned; whether this was because these were the only two covariates available for matching in three of the five target studies.

In terms of presentation, it is somewhat confusing that Van Sanden et al. (2016) tabulate the outcome of interest (Sustained Viral Response at 24 weeks) alongside the baseline patient characteristics without clear distinction, since it could be confused as a potential candidate covariate for matching. Additionally, in the final table reporting the results of the matching, compared with other publications, the presentation is less clear. It could be improved by reporting pre-matching and post-matching figures together and distinguishing baseline characteristics used for matching from the outcomes. This allows for quick verification of the success of matching and the impact on outcomes. As with most of the other articles reviewed, while Van Sanden et al. (2016) explain the limitations of their final analysis, for detail of the “MAIC covariate matching algorithm”, the reader is simply referred to (Signorovitch et al. 2010). The authors also employed less common terminology where they refer to the “pseudosample size”, “residual effective sample size”, “effective pseudo-population size” and “residual effective population size” instead of the more usual “effective sample size”. It is also stated in this article that “the general view in practice is that it should not be less than double digits” (Van Sanden et al. 2016). It is not clear where this rule comes from.

Another analysis by Van Sanden et al. (2018) made use of an unanchored MAIC to show the improvement in clinical outcomes using daratumumab as opposed to pomalidomide + low-dose dexamethasone for treating patients with heavily pre-treated and refractory multiple myeloma (Van Sanden et al. 2018). Presentation of the methodology is relatively clear although technical details are not discussed. Unlike the previous study with the same lead author, there are no details given on how the clinical trials were identified for the analysis. On the other hand, the selection procedure of matching covariates is straightforward. This article focuses more on the detail of comparing survival outcomes than in the previous application of MAIC, where time-to-event outcomes were also analysed.

The preliminary part of the analysis involved obtaining the individual patient time-to-event data for the comparator pomalidomide + low-dose dexamethasone using the “Guyot et al. algorithm”. Broadly speaking, this involves generating individual patient time-to-event data by using digitised Kaplan-Meier curves and published information on the number of events and number of patients at risk for various points in time (Guyot et al. 2012). Also in the article, an extension is presented in a sensitivity analysis where patients receiving daratumumab are restricted to those who were treatment naïve to pomalidomide in order to match those in the trials for pomalidomide + low-dose dexamethasone. This was not considered in the base case due to the ESS becoming too small after excluding the patients previously treated with pomalidomide. The base-case estimate of the improvement in the overall survival outcome for daratumumab is otherwise considered conservative, since a high percentage of the daratumumab patients were pomalidomide-treated and refractory. The limitations of their analysis are considered and described in some detail, including a sensitivity analysis where missing data on certain suspected prognostic variables was

imputed in order for those characteristics to be matched. Overall, while the description of the steps to the analysis are described clearly, there were some gaps, such as the lack of reference to the propensity score weighting approach that is central to MAIC. Confirmation of the methodology used to perform the adjustment would be of huge benefit to this study.

A recently published study showcased an extensive example of applying three MAICs. Maksymowych et al. (2018) were coherent in outlining the method while referencing the Signorovitch methodology and the NICE TSD. An interesting variant in this article is that multiple MAICs were required and reported as part of one analysis; owing to the multiple trials included. Two MAICs were developed between the competitor's trial (ATLAS – adalimumab vs placebo) and two individual trials from the drug company (MEASURE 1 and MEASURE 2 – secukinumab vs placebo). In addition, a third MAIC was performed which used a pooled dataset made up of those two trials (MEASURE 1&2) and these were compared against the competitor (ATLAS). Hence, a separate set of weights would be needed each for MEASURE 1, MEASURE 2 and MEASURE 1&2. The third analysis was presented because it resulted in a larger effective sample size. Another consideration in this article was the acknowledgement of imputed data, both in ATLAS and in MEASURE studies.

For comparisons made at weeks 8 and 12, all three MAICs could be anchored on a placebo arm (available in ATLAS, MEASURE 1, MEASURE 2 and hence MEASURE 1&2). Some participants in ATLAS who were administered placebo were then allowed to receive open-label treatment with adalimumab if they did not achieve at least a 20% improvement in ASAS response criteria at week 12. Thus, the comparisons beyond week 12 had to be unanchored. A similar procedure of offering active treatment was administered in MEASURE 1 and MEASURE 2 after week 16. Therefore, the unanchored comparison would use an adjusted secukinumab treatment response with adalimumab. The comparison was made for separate outcomes at weeks 16, 24 and 52. As noted in Section 3.3, the same set of patient weights can be used for multiple outcomes.

Although the presentation of results is comprehensive in (Maksymowych et al. 2018), once again there is little detail on the propensity score weighting approach. In fact, Maksymowych et al. (2018) state that MAIC “uses a form of propensity score weighting of individual patient data (IPD) to match patients from one trial with those from another for baseline characteristics, especially those that may influence treatment response”. This phrasing could give the impression that MAIC matches patients on a one-to-one basis, which is a misconception. While the article is thorough in its referencing, the complex trial design, imputation of missing data and variety of outcomes and comparators make the MAIC comparisons difficult to assess. Although the focus of this review is to identify understanding of the MAIC method within the literature, it is important to acknowledge the complexity to this analysis even prior to making use of MAIC.

In (Regnier et al. 2016) various methods for making an indirect comparison of ranibizumab and aflibercept are compared. These approaches depart from the methodology given in (Signorovitch et al. 2010). The term “MAIC” is not used in this paper, although one of the approaches for which results are presented is performed using similar principles. It is noted that while the Signorovitch approach was also applied in the course of conducting this study, the results derived were omitted because the standard deviations of the baseline characteristics could not be aligned to those of the competitor’s trial after weighting.

Odom et al. (2017) conducted an unanchored MAIC using IPD from a single-arm trial studying sonidegib and compared this against aggregate data for vismodegib for the treatment of locally advanced basal cell carcinoma. A naïve unadjusted comparison of the two treatments showed a longer progression-free survival and higher objective response rate for sonidegib, but the distributions of baseline characteristics that could be prognostic variables differed, hence confounding the treatment comparison (Odom et al. 2017). The process of identifying baseline characteristics suitable for matching is described clearly, as are the efforts made to align the two studies, e.g. in terms of outcome measurement. It is noted that matching variables were selected a priori and limited to two, due to the small sample size from the study for sonidegib.

Overall, the study by (Odom et al. 2017) is well described and the authors went further to explain the MAIC methodology than in other published applications, citing implementation of the Newton-Raphson algorithm in SAS to obtain the unique solution of $\hat{\beta}$ for the weights, as mentioned in Appendix B to (Signorovitch et al. 2010). Overall, the article closely follows the guidance issued in the Signorovitch article and TSD 18. An interesting aspect of this MAIC is that the treatment outcome results were largely unchanged from the unadjusted naïve comparison. Here the purpose of the MAIC was “to provide support for the validity of the naïve indirect comparison results” (Odom et al. 2017). This was the second motivation for MAIC identified in Section 3.3.

In (Warren et al. 2018), the article acknowledges that “although an indirect comparison is not a replacement for a head-to-head trial, accuracy is improved when a common active comparator is used with a method”. This mirrors a trend for the literature to implicitly refer to a hierarchy of evidence in terms of MAICs not being as robust as RCTs but carrying less assumptions than those required in an unadjusted indirect comparison. The authors make use of two weighting approaches which they term “SG total” and “SG separate”. Following correspondence with one of the authors of this article, it was confirmed that SG was an abbreviation for “Signorovitch”, although SG separate does not correspond to the methodology described by Signorovitch (2010). The descriptions provided in the article could include more precise detail, as the procedure for each is not fully clear. SG total is described as “calculating weights for the IPD based on mean values of covariates for total population” whereas SG separate bases the calculation “on mean values of covariates for each treatment arm separately”.

For this study, the analysts had access to IPD for several clinical trials examining ixekizumab against the different common comparators (IXORA-S with common comparator ustekinumab; UNCOVER-2 and UNCOVER-3 with both placebo and etanercept as common comparators and UNCOVER-1 as an additional source of information with etanercept as the common comparator). The analysts used AgD information from various clinical trials to inform the matching for each MAIC (CLEAR to base the MAIC anchored on ustekinumab, FIXTURE for the MAIC anchored on etanercept and the set of trials: ERASURE, FEATURE, FIXTURE and JUNCTURE for the MAIC anchored on placebo).

The term “total population” used in the definition for SG total is misleading, since the AgD is typically drawn from one trial, but it is likely this refers to the case where multiple trials were available for a single comparison between two treatments. This occurs in this analysis for the comparison between placebo and secukinumab, where data is drawn from the ERASURE, FEATURE, FIXTURE and JUNCTURE trials and combined. Since each trial would have its own set of patient baseline characteristics, a weighted average would have to be taken. Following clarification with one of the authors of this article, SG total is to be interpreted as following the usual anchored MAIC approach, where a single set of weights is generated based on the combined IPD set of control and intervention arm patients.

For SG separate, instead of calculating the set of weights in the usual manner using the information from the intervention arm and control arm together (in this case ixekizumab and placebo/etanercept/ustekinumab respectively), the selected patient characteristics of each arm are matched to their aggregate counterpart via a separate set of weights generated for each. This would result in six sets of weights in this example, with the baseline characteristics of ixekizumab patients matched to those of the secukinumab patients and each control arm being matched to their counterpart. Given the lack of studies which have made use of this modified method, it is difficult to establish what the implications are. However, it would not seem appropriate to treat the intervention and control arm differently through a separate set of weights.

Overall, the article by Warren et al. (2018) does raise the question of how data on multiple trials should be pooled for MAIC and considers whether an application of MAIC should use the method of separation or not. In this application, the impact on outcomes was relatively minimal as the outcome results did not appear to deviate hugely from those derived from the original method (Warren et al. 2018). Only the post-matching baseline characteristics that were obtained using the SG total method were reported in the table.

Several of the covariates that were matched in (Warren et al. 2018) were proportions and so the raw number of patients would not normally be reported after matching. However, in this article a new figure post-matching was reported based on multiplying the new proportion by the ESS and this is misleading. The ESS is an estimate as the weights are not fixed or known and so should not

be used to derive actual raw numbers of patients in this way (Phillippo et al. 2016). The ESS figures could also be better labelled as such, instead of being reported in the same row as the number of patients, N, in the table.

The authors report conducting a sensitivity analysis which exchanged two matching variables for two other characteristics yet in the table of post-matching values, it appears that all characteristics are matched. Unlike other publications, there was no table of outcome values provided. Overall, the presentation in (Maksymowych et al. 2018) is more precise by comparison and sets a more appropriate standard of reporting of MAIC analysis. Warren et al. (2018) mention use of the bootstrapping method to calculate the variance of the MAIC estimates. While this is accepted as an alternative method in the TSD, it does raise the question as to how MAIC results should be compared across appraisals if a methodological variant such as this is present, for reasons outlined in Section 3.5.1.

An example where a methodological variant is discussed in the TSD concerns an analysis by (Sikirica et al. 2013) which involved an unanchored MAIC despite a common placebo arm existing between the two treatments. Sikirica et al. (2013) justify the unanchored choice by constraining the weights such that the placebo arm outcomes were balanced for the two treatments. The authors of the TSD state that it is unlikely that this has the effect of randomisation and that this method variation needs to be evaluated more formally (Phillippo et al. 2016).

3.6.3 Summary of Review of MAIC Applications

The findings here motivate a more extensive review to investigate older literature involving “legacy” examples of MAIC published under different terminology before the more standard terminology arose. Overall, there were 16 applications of MAIC examined with publications dating from between 2010 and 2019.

It is also apparent that there may exist a wide variety of possible method variations in the literature. There was evidence of less common terminology, such as where Van Sanden et al. (2016) refer to the “pseudosample size” and “residual effective population size” among other terms instead of the effective sample size.

Some positive aspects of the literature included points where the implementation details were confirmed, such as use of particular statistical software, packages or clear reference to the methodology papers. As could be expected, there was a tendency for articles that were published more recently to articulate the MAIC method more effectively but there were exceptions. Publication of the TSD in 2016 would have almost certainly helped to consolidate the understanding of MAIC among investigators.

In summary, the majority of MAIC applications explored in this review would typically outline the problem of lack of head-to-head, randomised trials. This would be followed by an introduction to

the MAIC method as a means of matching available IPD to that of a competitor population and then reference would be made to propensity score weighting with estimation of weights via logistic regression. None of the reviewed applications gave further details on this key aspect of the methodology. There are two reasons for why this is important. Firstly, this is a critical aspect of the method. The re-weighting of individual patient data is central to MAIC and impacts on the results. With the current level of detail in the description of this process, it is difficult to envisage readers becoming fully informed as to how the weights are generated. Secondly, this process is subject to variation as identified in Section 3.5.2. This includes different weighting methods, as explored in (Petto et al. 2018) as well as the technique of weighting separately by arm identified in this review (Warren et al. 2018). There is also the question of which minimisation algorithm to use when calculating $\hat{\beta}$, the BFGS method is used in the R implementation produced in the TSD (Phillippo et al. 2016) whilst Newton-Raphson optimisation in SAS was proposed in (Signorovitch et al. 2010).

3.7 Existing R Packages

Using METACRAN, a search for R packages of relevance to MAIC or STC and had been uploaded to CRAN was carried out. As applied in the search protocol for published MAIC applications, a broad range of search terms was used including: “matching adjusted indirect comparison”, “maic”, “simulated treatment comparison”, “stc”, “population adjustment”, “matching”, “evidence synthesis”, “signorovitch” and “propensity score weighting” as well as basic variants on these search terms. Some relevant R packages were available on the Comprehensive R Archive Network (CRAN) at the time of this review. These included: `Matching`, `PSW`, `gemtc`, `pcnetmeta` and `netmeta`. A published review by (Neupane et al. 2014) compared the last three of these which all concern NMA. Their conclusions are summarised below.

The basic process for carrying out NMA begins with an exploration of the data on all competing treatments in their network form (the relationships that exist between the various treatments; whether they are connected or not). An assessment of the collection of studies in terms of heterogeneity and inconsistency is performed and any such inequalities are handled. A model is proposed and is assessed for fit using model diagnostics. The model will return estimates of relative effects between treatments or rank probabilities of treatments being the most effective. Both `gemtc` and `pcnetmeta` operate according to the Bayesian framework while `netmeta` models under the frequentist framework, however both frameworks may lead to generation of treatment ranking probabilities. Only binary outcomes can be handled in `pcnetmeta`, whereas `gemtc` and `netmeta` may also handle count, continuous and survival outcomes. Neupane et al. (2014) continue to compare and contrast the number of modelling features that the three packages provide. They conclude that `pcnetmeta` is less comprehensive in this regard, with fewer descriptive measures although a number of plotting options are available. The focus in `pcnetmeta` is on simplicity and ease of use which is a key aspect to any R package. The `netmeta` package does not provide the

Akaike Information Criterion (AIC) for assessing goodness of fit for the NMA model whereas both of the Bayesian setting packages include the Deviance Information Criterion (DIC). While these packages are not directly related to implementation of MAIC, it is useful to review the choice of packages available for conducting NMA when considering the incorporation of MAIC results into a network meta-analysis. Neupane et al. (2014) conclude that between the three packages, they offer nearly all the important functionalities for conducting NMA.

The `Matching` package is described in (Sekhon 2011) and is concerned with implementing a variety of multivariate matching algorithms. The package contains a function with a choice of algorithms: propensity score matching, Mahalanobis, inverse variance and genetic matching. A separate function assesses covariate balance. Overall, the package provides a template for how variations of matching techniques could be implemented in a single package and acts as an educational tool for the topic of matching for causal inference. `PSW` is a simple package which offers propensity score weighting methods including inverse probability weighting for estimating average treatment effect. However, the package is basic and does not accommodate indirect comparisons. More importantly, the method for calculating the weights is not set up for user modification and little documentation from the package is available for determining the method adopted. As with the `Matching` package, the `PSW` package is concerned with propensity score methods which are concerned with estimating outcomes in a target population based on a sample subpopulation which differs in terms of covariate distribution.

None of these packages are designed specifically to implement the inverse propensity score weighting for MAIC or the outcome regression modelling for STC.

A similar search was conducted on Github with the expectation of identifying code development related to or intended for implementation of MAIC. Github can be used in conjunction with RStudio as a means of version control and the website hosts a diverse range of collaborative software development projects using a multitude of programming languages. This form of searching was challenging due to the sheer volume and variety of projects uploaded on Github, many of which were incomplete or did not represent fully working R projects. In spite of this, there were no development files or projects identified that pertained to a working implementation of the method.

While not an R package, one piece of useful software uncovered during the search was an Excel-based tool which performs the indirect comparison, as per the Bucher method (Bucher et al. 1997). The tool is implemented in Visual Basic and as of the time of review, is available from the website of the organisation which developed the tool, the Canadian HTA body (Wells et al. 2009).

3.8 Overall Literature Review Summary

The findings of this review are that MAIC remains a relatively new methodology. Further publication of the methodological detail would be of interest since, to some extent, it has been challenging to

identify the commonly accepted definition of MAIC. As recent as 2019, in a review of technology appraisals to NICE, it was reiterated by authors of TSD 18 that research to improve on the method was needed (Phillippo et al. 2019).

The current landscape of MAIC applications in the literature feature a wide variety of naming conventions, use of alternative approaches to Signorovitch et al. (2010) despite reference being made to “matching-adjustment”, and a distinct lack of detail on the weighting approach used. Examination of existing applications of MAIC showed that discussion of the methodology was lacking. This results in publication of adjusted results of treatment effect without sufficiently clear derivation. On the other hand, one strength of the literature is that acknowledgement is generally (though not always) made of the various limitations that exist in their analyses, such as exclusion of potentially relevant trials, reduction of effective sample size and the fallibility of indirect comparisons such as effect modifiers that remain unaccounted for. The TSD also draws the conclusion that there is a need to standardise the MAIC method and this idea is the central motivation for this research project. Finally, the search for existing R packages in this review showed that at the time of writing, no implementation of the method was available.

Together, these findings motivate the creation of a package in R which will standardise the MAIC implementation, both in terms of following the procedures and selection of parameters in line with the core literature. The package itself, once adequately designed, can contribute to standardising the implementation as a result of the traction gained by investigators who use the package for their analyses.

Chapter 4 R Package

Following the literature review of MAIC, this chapter describes the R package which was built with the aim of facilitating straightforward implementation of the method. In general terms, the R package is designed to equip the user with a set of functions which will enable them to input the various pieces of data necessary for an anchored MAIC. This comprises the IPD and treatment outcome data. The user will then be able to perform the calculations and view the output (including weights applied to the IPD, analysis of the re-weighting and the resulting population-adjusted treatment effect).

Section 4.1 reiterates the aim of developing the R package and outlines the motivation for choosing the R programming language. The purpose of developing R packages generally and the software development workflow adopted in the building of `maicer` is presented in Section 4.2. Some requirements for the package specified prior to development are outlined in Section 4.3 and this is followed by the detail of the development process in Section 4.4. The chapter concludes with a walkthrough of the package showcasing its functionality with screenshots of the RStudio console in Section 4.5.

4.1 Motivation

It is expected that the establishment and free publication of this standardised tool on (GitHub 2019) and CRAN would allow interested HTA stakeholders with access to their own IPD to carry out their own MAIC. Additionally, they will be better equipped to interpret a pre-existing MAIC that was built using this package, with the knowledge that this has been carried out according to a specific range of parameters. The motivation for this is that a published MAIC would be reproducible, were the user to have access to the same IPD. Furthermore, given the highly sensitive nature of IPD making it difficult to access freely, having the assurance that a given MAIC has been conducted according to a standardised tool is of value.

The R programming language is an appropriate platform for this tool for several reasons. Firstly, it is free and open source. Given the vast number of potential stakeholders interested in a standardised tool for MAIC, having an open-source environment is of value for disseminating the tool and stands to benefit from a wider uptake. Secondly, R has been used for cost-effectiveness models in HTA previously and continues to be used, notably in cases where calculations are too complex for other software to handle sufficiently. The release of a purpose-built R package for health economic evaluation further supports this (Filipovic-Pierucci et al. 2017). Finally, the example R code provided in Appendix D of the TSD (Phillippo et al. 2016) provided the most comprehensive overview of the method implementation and the document remains a key point of reference for this topic.

4.2 Package Development Workflow

In R, a package is the key template for making code easily tested, generalised, documented and sharable (Wickham 2015). A package would perhaps more accurately be termed as an extension, as indicated in the official development manual (R Core Team 2019), since it is a means for the user to access additional functions in the R environment. The package format proves to be invaluable in terms of standardising the way in which R users access these extensions built by other users. The standard package format also equips all R package developers with a standard convention for organising their work.

A number of good practices are identified by (Wickham 2015) in terms of R package development; these were adopted insofar as was possible for the development of this package. A key guideline is the importance of iterating development from an early stage and regularly. In other words, once a new piece of functioning code has been implemented, it should be run and used for its intended purpose. These steps should be repeated for each new piece of functionality. The package was developed in RStudio in a Project file format. RStudio Project files are useful as they allow packages under development to be accessed in isolation (away from other work in RStudio) under their own working directory. This becomes useful for testing the package in a separate, clear RStudio session, as if the package was just installed. It also makes it clear to the developer the separation between the package contents and other R script files. RStudio Project files also facilitate version control via Git/GitHub.

The process of package development is helped through the use of the R package `devtools`, which simplifies development by providing the user with a set of functions which will automatically perform the desired development actions. For example: `devtools::create()` will create a new package. A standard R package is composed of seven key parts which are described in Table 3 below:

Table 3 The seven common components to a standard R package

DESCRIPTION	The DESCRIPTION file provides a standard set of key information on the package including version, the package creators, package dependencies, licensing and more.
R/	The R/ folder houses the R code; this is the barebones of the package.
tests/	The tests/ folder helps the developer to implement a set of tests from a single location. It is important to iteratively run newly implemented code as part of package development. The tests/ folder facilitates testing from a package overview perspective such that if tests become redundant, they may be updated as the package is developed. This is often referred to as unit testing,

	<p>which is a manner of structured testing such that each component of a software behaves as it should do, with the appropriate inputs and outputs. The package testthat helps to facilitate testing for the developer by automatically assembling the tests/ folder.</p>
man/	<p>The man/ folder holds all the package documentation i.e. descriptions of package functions available to the user and examples of their use. This can be essential for sharing a package so that users may query package functions and find out what they do. The package roxygen2 simplifies the documentation process by generating the .Rd files for functions based on comments the developer enters alongside their functions they wish to document. Vignettes and package-level documentation may be implemented using R Markdown; a system of tags which aid in the process of producing a document as with other mark-up languages. The man/ folder may also document datasets stored in the data/ folder.</p>
data/	<p>Datasets may be stored in the data/ folder which can be useful for providing prospective package users with a bespoke dataset that showcases the package functionality, or simply in order to help them understand how to use the package through the use of examples.</p>
vignettes/	<p>The purpose of a vignettes/ folder is to store documentation in a longer form that provides a general overview of the package and the broader problem it aims to solve. This can prove useful in cases where the package is facilitating a complex statistical algorithm, where the finer details of the calculations may need to be explained.</p>
NAMESPACE	<p>The final common element to a package structure in R is the NAMESPACE file. This file defines how the package should interact with other packages in terms of imports and exports. Imports refer to functions from other packages that are used by the package under development. Exports refer to functions from the package being developed where it is proposed that they be made accessible to the user once they load the package. In other words, it is preferable to “hide” functions that run in the background and are not required by the user to perform the desired actions. The NAMESPACE file defines these two distinctive groups. Whereas the DESCRIPTION file will define what other packages are required, the NAMESPACE file refers to specific functions from other packages.</p>

For large-scale projects in particular, version control can prove invaluable to development. Git is a system which enables tracking of changes to code and can be connected to GitHub; a website for sharing and hosting code that is particularly popular among R package developers. This version

control system can interface with an RStudio project file for seamless use and GitHub allows for collaborative work on fixing bugs and improving packages further.

Anyone may share code on GitHub, however, once a package is ready to be shared more formally, it can be submitted to the Comprehensive R Archive Network (CRAN). Owing to the network being managed, there is a package submission process to be followed. An automated R CMD Check must be run and passed as part of this. The automated check is in fact a series of tests covering a variety of common problems, mostly concerning package structure. The package can continue to be maintained and updated with patch releases following successful acceptance onto CRAN.

4.3 Requirements

The key requirement for the package was that it would aid the user in being able to conduct MAIC with ease. From the outset, the scope was left relatively broad so as to determine the extent of programming required and to allow flexibility for the scope of the package to change. The basis for the package would be to generalise and extend the example implementation provided in Appendix D of the TSD (Phillippo et al. 2016) as this was the most comprehensive overview identified from the literature review. The appendix relates to an anchored comparison which is arguably less straightforward to implement as trial arm data must be taken into consideration. The package was built for the anchored scenario in the first instance but allows the user to calculate the set of weights needed for an unanchored scenario. The user would then need to apply the weights to the IPD and using the weighted outcomes calculate the new summary outcome before conducting the indirect comparison with the comparator of interest. It would be possible to format the package to allow for a more user-friendly implementation of the unanchored scenario. By the end of this project, the package was instead structured such that the user may work with one of two object formats: one which performs the adjustment to estimate the weights (and output these) and another which also performs the subsequent indirect comparison for MAIC. The details of this structure are described in Section 4.4.3. The package implementation hereby refers to the anchored scenario unless where stated otherwise.

Development of the package homed in on three key aims: scope, generality and usability. These are defined here for the purposes of this project. The first focuses on maximising the functionality of the package so that it may be able to cater for different variations of the MAIC method, as explored in the literature review. Ultimately, a limited number of these were built into the final package during the course of this project. Generality refers to the ability of the package to interpret trivially different forms of input from the user and reduce the need for defining all implementation requirements in detail according to a specific template. In this case, some of the functions were adapted to accommodate varying amounts of user input. Finally, usability refers to the ease with which the tool may be used. The package functions were refined so as to require a minimal amount of code executions from the user in accordance with the philosophy of R package developer,

Hadley Wickham, in his book *R Packages* (Wickham 2015). This simply states that anything that can be automated, should be and as much as possible should be performed by functions. The ultimate goal here would be to develop a user-friendly point-and-click interface which would overlay the package functions. This would negate the need for the user to have to execute their own code directly into the R console. This proposal, among others for extending the work done, is outlined in Section 5.1.

4.4 Development

The appendix to the Technical Support Document outlines an implementation for matching one continuous variable to a mean and standard deviation statistic. The resulting estimated weights are then used to adjust a binary outcome and perform the indirect comparison. Following this example, development of the R package, called *maicer*, began with the aim of accepting a given set of IPD and to perform the matching for one continuous variable. The target summary data in this instance would be a mean statistic. The code was adapted to allow for matching to a standard deviation. Since standard deviation is a function of variance, this involved incorporating the second moment of the covariate into the centring of the matrix of effect modifiers, as described in Section 3.5. The next step was to extend this to cater for matching multiple variables of a continuous type simultaneously. Subsequently, the code was adapted to facilitate dichotomous variables. The literature review of published MAIC analyses revealed a trend of these “proportion-type” variables being commonly reported and used for matching. This type of covariate can be recorded within a set of IPD as a binary variable and treated as such by the R package.

Further development was geared towards accommodating matching to a median statistic. However, preliminary testing of this functionality in the prototype suggested that the resulting weights were highly skewed. There were also challenges in determining if the median of the weighted IPD would reflect the target median entered. Since both the literature review and consultation with industry experts indicated that the median was less commonly reported in aggregate summary data in general, this functionality would be less essential and was omitted for the first build. The problems encountered are explored further and reported in Section 4.4.2.

Once the package could generate the desired weights and carry out the matching-adjustment, for the final anchored comparison, it would also need to carry out the indirect comparison, for which further user input would be necessary. Using the binary outcome case, the package was extended to accept patient outcome and trial arm data, before generating the final population-adjusted relative treatment effect between the intervention treatment (for which IPD is available) and the comparator treatment (where only summary data is available). The standard error for the estimate was also calculated. For this indirect comparison, information on the relative effect between the comparator treatment and the anchor treatment would also be needed as illustrated in Equation 4, reproduced below for ease of reference:

$$\hat{\Delta}_{BCAC} = g(\bar{Y}_{C(AC)}) - g(\bar{Y}_{A(AC)}) - (g(\hat{Y}_{B(AC)}) - g(\hat{Y}_{A(AC)})) \quad (4)$$

With a multitude of clinical study designs and outcome measurements, it would be best to follow the generalised framework published by (Dias et al. 2013, Dias et al. 2011b) to ensure different outcome types could be incorporated. Expanding the tool to do so would be a desirable goal, however due caution would be advised for ensuring that the appropriate comparison is being made and that the values are being interpreted correctly. The importance of scale was explored heavily in the Technical Support Document (Phillippo et al. 2016) and so the indirect comparisons tool produced by CADTH (Wells et al. 2009) may act as a useful complement to the weighting provided by this R package, as described in Section 5.1.

4.4.1 Object Orientation

The penultimate stage of package development was a restructuring of the code so that it would be object-orientated and easily sharable in accordance with the approach published in (Leisch 2008, Tierney 2016). The official manual on writing R extensions made available from the R Development Core Team was also consulted (R Core Team 2019). The purpose of object-orientation here was to reduce the user's reliance on numerous functions to carry out the analysis they require. Instead, they would only have to execute a minimal number of lines of code to perform the desired calculations, thereby meeting the original objective of usability for the package. By contrast, a functional programming structure would typically require the user to remember how to use a number of functions, their behaviour and the correct order in which they should be used. In object-orientation, an object is a data structure with a set of attributes. Classes define how objects should be represented (like a blueprint) and methods define specific actions to be performed on objects of a particular class. An object is simply an instance of a class as called upon by the user; the object stores the results obtained when the class is used for a given dataset (Leisch 2008).

In R, three object systems are available: S3, S4 and RC (Reference Classes) (R Core Team 2018). RC most closely resembles other typical object-oriented programming languages such as Java. These typically require a method to be sent to objects and the object determines how to handle the method. S3 is a simplified system that does not formally define classes, the class can just be simply assigned to an object. On occasion, S3 leads to ambiguity if a function that carries multiple definitions is used on an object of a particular class; necessitating the user to use the double colon operator. For example, this could arise if the function `plot()` was defined differently across multiple R packages and these were being used in the same piece of code. S4 is a more formal system than S3 and requires definition of the class. This can be important for larger scale projects where more planning is required. S3 is the system of choice for much of R programming and makes the coding language more accessible to users of different backgrounds. The RC system would be more akin to the object-oriented programming environments that experienced developers may be more familiar with. S3 methods are called in a different way. S3 does not formally define classes,

the class can just be simply assigned to an object. In the `maicer` package the S3 system was used as there was no obvious need for a formally defined protocol as in S4. In many projects the choice of S3 is straightforward. Building S3 Methods allows for functions to be generalised so as to behave in a specific manner according to a defined protocol. A requirement of building one's own class is to define behaviour for three rudimentary functions referred to as *generic functions*. These are: `print()`, `summary()` and `plot()`, though this last function is not strictly necessary. `print()` defines the behaviour of the object when it is called directly from the console. It should detail the key results held within the object. There is no specific definition for the `summary()` function, though it should produce summary statistics and some more detail held within the object.

4.4.2 Challenges in Matching to a Median Statistic

Various published MAICs attempt matching a covariate to a median statistic instead of or as well as a mean, depending on the AgD available (Signorovitch et al. 2013, Tremblay et al. 2018). Initially, a median statistic may appear to be a superior summary statistic to use for matching; it is not impacted by outliers and in the case of discrete data, will take on one of the discrete values in the set when the sample size is odd. The median can be regarded as a suitable choice for data that does not follow a normal distribution. However, the methodology for MAIC described by Signorovitch et al. (2010) relies on the method of moments to match the available IPD (treated as a sample) to the published AgD (in effect treated as the theoretical moment or as the parameters defining the population).

There is scope for debate on the use of the method of moments as a means of performing matching-adjustment and this may have given rise to proposals for other methodologies as identified in the literature review. Based on the experience in developing the R package and subsequent testing on some simulated clinical trial data, simulation studies investigating the behaviour of resulting distributions from different sets of weighted IPD output in MAIC would be of interest. Nonetheless, the debate on the appropriateness of the methodology was not the focus of this research project. For the purposes of this thesis, it is assumed that for the indirect comparison scenario previously outlined, matching moments to achieve population adjustment is a suitable method.

Returning to the choice of matching to a mean or median statistic, the statistical properties of the mean prove more useful with respect to the method of moments. The definition of the first moment has the same definition as the formula to calculate the sample mean. It is therefore simple to verify that a mean statistic has been matched. By contrast, the median itself carries little distributional information. In reality, range and other quantile information would also be required in order to perform a reasonable degree of matching. Sometimes, range information can be abstracted from inclusion/exclusion criteria of patients reported from the relevant clinical trial, as per the methodology described in Section 3.5.

Even when adequate distributional information is available, there remains the problem of verifying successful matching following the reweighting of the data. In the first published MAIC article, Signorovitch et al. (2010) outline that reported medians (or other percentiles) can be matched by converting a continuous covariate into a binary characteristic with a True/False value denoting whether each patient in the set of IPD is greater than the specified median (Signorovitch et al. 2010). The proportion of cases in agreement can be derived from this and this proportion can be matched to equal 50% as with any binary characteristic. In reality, while the mean of the binary characteristic is being matched, conceptually, it is not clear how the individual ages of the patients will be represented in the weighted set of data. In essence, the weighted data must result in the value for the binary characteristic switching for a number of patients so that the two groups are even, but it is not known whether the switched patients should take the median value or how far their value should extend into the opposite group. As an example, if the patient age of the IPD set needs to be matched to a median of 40 and the original median is 37, then a number of the patients need to be weighted so that they assume a value equal to 40 or above. Unlike the mean, there is not enough distributional information to determine how this assignment should work in practice. Relative to the median statistic, there is not much to distinguish other patients except for being above and below the median.

Problems can also arise where many patients correspond to the old median value, particularly in the case where the sample size is small. For this scenario, it can be difficult to determine how these values should be adjusted; whether the weighted value should cause them all to be randomly assigned above and below the new median. Additionally, MAIC works optimally for the case where there is good patient population overlap and the ranges of the AgD summary statistics are a subset or similar to the corresponding values for the IPD (Signorovitch et al. 2012). In the case of a supplied median and range where the range is slightly wider than that of the IPD, it is not clear how it would be possible to weight the IPD such that it would have a range matching the AgD where the AgD range is slightly wider. This could be characterised as a classification problem. Overall, while some of these challenges may not present a huge difference to the end MAIC estimates, it is plausible that in some situations, such as small patient sample sizes, there would be a notable impact on the results. Research into these questions would be of interest.

4.4.3 Design

A simulated individual-level patient dataset was generated using the `wakefield` package in order to drive development. This dataset consisted of 200 patients and was extended to one million patients for testing purposes (naturally, a patient dataset of this size would be unrealistic). The simulated dataset from the TSD appendix was also recreated in order to test the package and as a verification exercise. The TSD dataset and the simulated set of 200 patients have been included in the `data/` folder with example code provided for the user to recreate the analyses performed. A walkthrough of the package with screenshots is provided in Section 4.5 of this thesis.

The design of the R package was formulated in light of the commentary from the TSD. The importance of the selection of covariates to match on (rather than matching on all covariates available) was reiterated throughout the document. The package therefore requires the user to set up a `matching_set` of covariates; a subset of the IPD. The set of covariates is initialised via the function `set_matching()` with arguments taking: the name of the IPD set, the name of the first covariate to match on as well as the target summary data for the respective covariate. If a standard deviation is entered, then the covariate's second moment is automatically appended to the set (as this is a requirement for the subsequent weight estimation). The user may then expand the set via `add_matching_covariate()`, start the process over again with `set_matching()`, or remove an individual covariate using `remove_matching_covariate()`. These functions build a `matching_information` list object which consists of the `matching_set`, the target information entered and a calculation of the relevant summary statistic information prior to the MAIC weighting.

In this way, the user is forced to first consider which covariates they wish to use for matching. In the initial design the user was required to specify the target summary data in a separate vector form and ensure that this was in the same order as the order of the covariates being entered. A reorganisation of the code improved on this so that specific target summary values could be entered for each covariate; reducing the chance of programming error while building the set of covariates to match on.

Once the set of covariates to match on has been specified, the user may obtain the estimated weights that would be applied to the IPD for MAIC. The functions that perform this calculation are not intended to be directly available to the user as they are internal, but they may be accessed via the triple colon operator. With the weights generated, the user is equipped to calculate the population-adjusted treatment effect for the trial where IPD is available. The weights can be applied to the original trial outcomes in the IPD set to give the outcomes adjusted for the other trial population. However, no information on the population-adjusted relative treatment effect can be generated without further information on the existing, unadjusted relative treatment effect between the comparator and anchor treatment.

In accordance with the framework described above, the final design led to the creation of two package objects. The first of these is a `match_adjustment` object, used for estimation of weights. The object may also accept outcome and trial arm information in order to illustrate the impact of these weights on the patient outcomes and hence calculate the population-adjusted treatment effect. For the final indirect comparison for MAIC, a second object termed `maic` is provided. Once again, optional arguments are provided for this object in order to accommodate varying amounts of information from the user. Both of these objects have separately defined generic methods and are documented as such. It is anticipated that prospective users may have varying levels of information available to them or may by choice prefer to simply obtain the weights for matching

adjustment. This package aims to be flexible by accommodating both levels of information and provide functionality for both.

The objects and functions available to the user are summarised in Table 4 below:

Table 4 Summary of objects and functions in R package

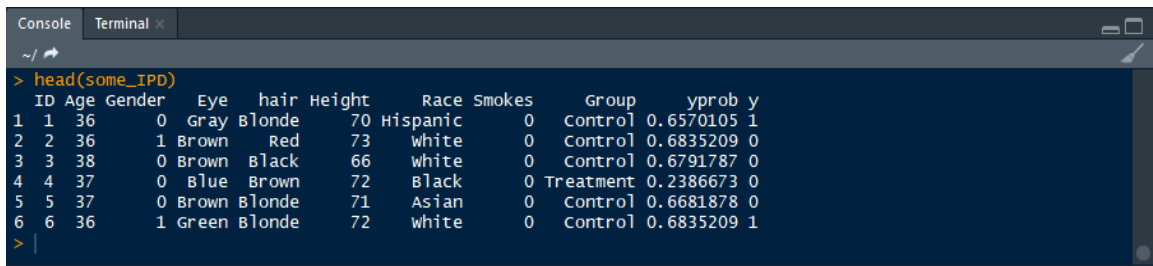
Functions	Explanation
<code>set_matching()</code>	Create a <code>matching_information</code> object which will store the covariates selected for matching in a <code>matching_set</code> object. A covariate must be entered to initialise the object as well as the target summary data to match on for this covariate.
<code>add_matching_covariate()</code>	Add another covariate to the <code>matching_set</code> object. The arguments for this function require the <code>matching_set</code> to modify, the set of IPD to draw the new covariate from, the covariate identifier and the target summary data to match on for this covariate.
<code>remove_matching_covariate()</code>	Remove a single covariate from the <code>matching_set</code> object.
Objects	Explanation
<code>matching_information</code>	A list-type object generated from the <code>set_matching()</code> function which stores all information entered prior to matching including: <ul style="list-style-type: none"> - the <code>matching_set</code> - the target summary-level data to match on - the values of the corresponding summary-level data in the <code>matching_set</code> prior to matching. <p>This object is created in the background and the user does not have to handle this object type.</p>
<code>matching_set</code>	This object is stored as part of the <code>matching_information</code> list and stores the set of covariates from the IPD set to be used

	for matching and is therefore used in the calculation of the weights for MAIC. This forms part of the <code>matching_information</code> object which is created in the background and does not have to be handled by the user.
<code>match_adjustment</code>	One of two possible objects which stores the output of the package. The <code>match_adjustment</code> object will store the weights and ESS for the MAIC but will not perform the indirect comparison.
<code>maic</code>	The second of two possible objects which stores the output of the package. A requirement of the <code>maic</code> object is information on the patient outcomes as well as relative effectiveness information between the anchor treatment and comparator treatment. This will return the final MAIC estimates.

4.5 Package Walkthrough

An example output from the package is presented in this section as a walkthrough using multiple screenshots from the RStudio console. The walkthrough illustrates how a given set of IPD prepared for MAIC analysis can be passed into a series of functions to return MAIC results. Results from the analysis are presented at the end showing the indirect comparison between fictional drugs “loloplitin” and “sonaliptin”. The MAIC is anchored on fictional common comparator “trizepibronate”. Note that the walkthrough and screenshots presented below relate to the initial `maicer` package version 0.0.0.9000. Functionality is currently limited to binary patient outcomes only.

Load RStudio and ensure the `maicer` package is installed and loaded. The following walkthrough illustrates an implementation of the package using the `some_IPD` set available in the `data/` folder.



```
> head(some_IPD)
  ID Age Gender Eye hair Height Race Smokes Group yprob y
1  1  36     0 Gray  Blonde   70 Hispanic     0 Control 0.6570105 1
2  2  36     1 Brown   Red    73 white       0 Control 0.6835209 0
3  3  38     0 Brown  Black   66 white       0 Control 0.6791787 0
4  4  37     0 Blue   Brown   72 Black       0 Treatment 0.2386673 0
5  5  37     0 Brown  Blonde  71 Asian      0 Control 0.6681878 0
6  6  36     1 Green  Blonde  72 white       0 Control 0.6835209 1
```

This is a rudimentary set of IPD generated using the `wakefield` package. Each patient is attributed with:

- An ID number
- An age ranging between 35 and 45
- A gender coded as a binary {0,1} variable. 54% of the patient set is classed as female.
- An eye colour, recorded as a categorical variable which can take one of five values: Gray, Brown, Blue, Green or Hazel.
- A hair colour, recorded as a categorical variable which can take one of four values: Black, Blonde, Brown or Red.
- A height ranging between 56 and 77 inches.
- A race, recorded as a categorical variable which can take one of six values: Asian, Bi-Racial, Black, Hispanic, Native or White.
- A smokes variable coded as a binary {0,1} variable. 17% of the patient set is recorded as a smoker.
- A dichotomous variable indicating whether the patient was assigned to a control arm or the intervention arm, coded as “Control” and “Treatment” respectively.
- A randomly generated probability of the patient experiencing the binary event of interest.
- A binary variable describing whether the patient experienced the binary event of interest (coded as {0,1}).

Note that proper internal correlation within the individual patient data is not reflected (such as appropriate distributions of height for males and females to reflect general population norms), however, the IPD and treatments concerned here are fictional and the results themselves are not intended to be meaningful.

We begin by specifying the covariates from the set of IPD that we wish to match on. Using the `set_matching()` function, the matching set may be initialised. Here the age covariate has been selected to match to a target mean of 42 and a standard deviation of 3.082363 (the more precise specification of standard deviation to more decimal places is arbitrary and has been done for illustrative purposes):

```
Console Terminal x
~/
> covariates_to_match <- set_matching(some_IPD, "Age", target_mean = 42, target_sd = 3.082363)
> covariates_to_match
$matching_set
  Age Age^2
1  36  1296
2  36  1296
3  38  1444
4  37  1369
5  37  1369
6  36  1296
```

It is not necessary to inspect the `matching_information` object created by the function as above; the first element of the list that is created will contain the full set of patients in the dataset (the `matching_set` object). The screenshot below shows the remaining elements of the list that have been assembled from initialising the set which include the target values of the AgD and the corresponding values of these summary statistics in the matching set prior to matching:

```
Console Terminal x
~/
$mean_statistic
  Age
39.975

$targets
  Age.mean  Age.sd
42.000000  3.082363

$sd_statistic
  Age
3.269222

> |
```


Two more covariates are selected for matching (Height and Smokes), with target summary statistics to match to (target mean height of 67 inches, standard deviation 3 and a target of 22% of the patient set to be recorded as a smoker):

```
Console Terminal x
~/
> covariates_to_match <- add_matching_covariate(covariates_to_match, some_IPD, "Height", target_mean = 67, target_sd = 3)
> covariates_to_match <- add_matching_covariate(covariates_to_match, some_IPD, "Smokes", target_proportion = 0.22)
> covariates_to_match
$matching_set
  Age Age^2 Height Height^2 Smokes
1  36 1296   70   4900      0
2  36 1296   73   5329      0
3  38 1444   66   4356      0
4  37 1369   72   5184      0
5  37 1369   71   5041      0
6  36 1296   72   5184      0
7  39 1521   70   4900      0
```

If we wanted to verify what summary data was entered, we can call the \$targets element of the object:

```
Console Terminal x
~/
42.000000  3.082363  67.000000  3.000000  0.220000
$sd_statistic
  Age Height
3.269222 3.523018
$proportion
Smokes
0.19
> covariates_to_match$targets
  Age.mean Age.sd Height.mean Height.sd Smokes.proportion
42.000000  3.082363  67.000000  3.000000  0.220000
> |
```

With a set of covariates selected for matching and the target summary statistics entered corresponding to each covariate, we are ready to find the estimated weights to be applied to the IPD for matching-adjustment. Applying the function `match_adjustment()` to the set of covariates will return the following output:

```

Console Terminal x
~/
> weighting <- match_adjustment(covariates_to_match)
> weighting
weighting IPD for MAIC resulted in the following:
ESS: 95.27659 (ESS/N): 0.476383

Summary of weights generated, scaled to original weight:
(1 = no change, <1 = downweighted, >1 = upweighted)

      v1
Min.   :0.02081
1st Qu.:0.34025
Median :0.68355
Mean   :1.00000
3rd Qu.:1.18937
Max.   :5.53786

The rescaled weights can be accessed by calling $rescaled from the match_adjustment x.
See summary() for more information. See plot() for a histogram of the rescaled weights.

 [1] "beta"           "centred"         "ess"
 [4] "essprop"        "match_adjusted_covariates" "matching_set"
 [7] "rescaled"       "rescaled_weights_hist"  "targets"
[10] "verification"  "verification_plot"    "weights"

For verification of successful matching, see $verification and $verification_plot.
> |

```

This provides information on the distribution of the weights, the ESS resulting from applying the weights to the IPD and the ESS as a proportion of the original sample size (ESS divided by the original IPD size of 200). In this example, no patient outcomes or arm identifier information were passed into the `match_adjustment` function so no further information on the resulting patient outcomes following matching-adjustment is available.

Further details of the output can be obtained using the `summary()` function. For this example where patient outcomes were not specified, the only output was the generated set of weights. The `summary()` function returns some details on the properties of the set of weights. This includes an excerpt of the centred covariate matrix (first 6 rows) to illustrate what was used for calculation of the weights. This is followed by a description of the basic distributional properties of the set of weights which are termed as 'rescaled' as they are divided by the total sum of the weights applied. Therefore, the interpretation of a weight less than one is that the individual patient has been weighted down, greater than one and the individual patient has been weighted up. The ESS and the ESS as a proportion of the original sample size are also presented in this output. The final component to the summary details how closely the weighted data matched to the targets entered. This is presented as a table. In this example each target summary statistic is listed with the target value, the value of the summary statistic in the weighted data and a measurement of the difference which has been termed here as 'divergence'. The percent divergence indicates the difference as a

proportion of the target value, represented in percentage form. As the percentage difference (indicated as percent_divergence below) is low (below 0.005% difference), this indicates close matching.

The output below illustrates that the difference between the desired target values and the actual summary statistic values of the weighted IPD are very small:

```
> summary(weighting)
weighting IPD for MAIC resulted in the following:
Covariates used in matching (centred):
  Age   Age^2 Height Height^2 Smokes
1  -6 -477.501    3    402  -0.22
2  -6 -477.501    6    831  -0.22
3  -4 -329.501   -1   -142  -0.22
4  -5 -404.501    5    686  -0.22
5  -5 -404.501    4    543  -0.22
6  -6 -477.501    5    686  -0.22
(Showing up to the first 6 patients)

summary of weights generated, scaled to original weight:
(1 = no change, <1 = downweighted, >1 = upweighted)

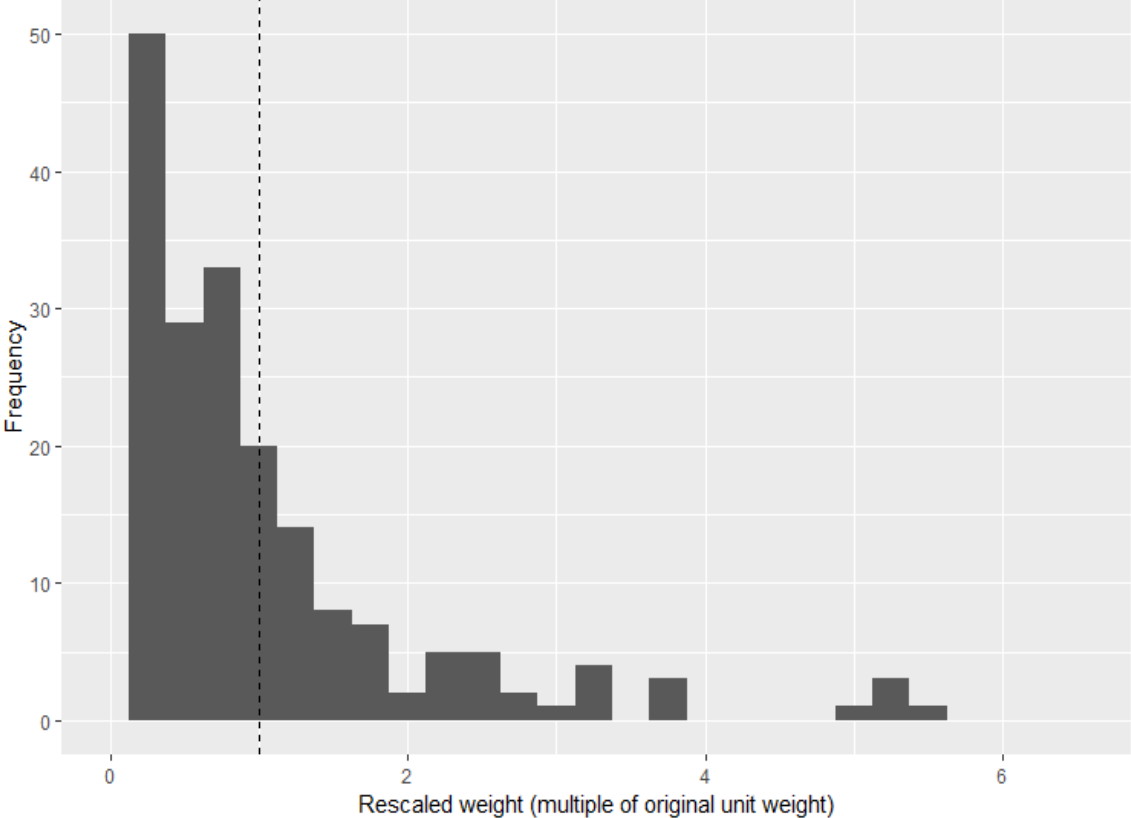
      v1
Min.   :0.02081
1st Qu.:0.34025
Median :0.68355
Mean   :1.00000
3rd Qu.:1.18937
Max.   :5.53786

The Effective Sample Size (ESS) is:
95.27659

The ESS as a proportion of the original sample size is:
0.476383

Verification of newly weighted IPD summary statistics matching the target AgD:
summary  targets weighted_data  divergence percent_divergence
Age.mean  Age.mean 42.000000  42.0000323  3.227667e-05  8e-05%
Age.sd    Age.sd   3.082363   3.0823312  -3.178192e-05  -0.00103%
Height.mean Height.mean 67.000000  66.9999926  -7.425819e-06  -1e-05%
Height.sd Height.sd  3.000000   3.0000102  1.022826e-05  0.00034%
Smokes.proportion Smokes.proportion 0.220000  0.2199992  -8.091784e-07  -0.00037%
> |
```

The distribution of the rescaled weights can be plotted by using the `plot()` function on the `match_adjustment` object. As previously stated, if all rescaled weights were equal to 1 this would indicate no change to the IPD, and we would expect the ESS to equal the original sample size (no adjustment would have taken place). In the histogram below the rescaled weights have been graphed within bins of 0.25. The vertical dashed line indicates the area of the graph corresponding to the number of weights within a 0.25 bound of 1 (between 0.875 and 1.125). Weights below the value of 1 have reduced the weight of an individual patient while weights above 1 have increased the weight of an individual patient:



Viewing a histogram of the weights such as in the plot above can be useful for gaining more information beyond the ESS value. It can help determine if any individual patients are contributing disproportionately to an adjusted analysis and enables the analyst to visualise the degree of distortion occurring in applying weights for the MAIC.

If the user has access to the trial arm identifier and patient outcomes, these may be passed into the `match_adjustment()` function in order to calculate the impact on the average outcomes by treatment arm before and after weighting according to population adjustment:

```
Console Terminal
~/
> matched_with_outcomes
weighting IPD for MAIC resulted in the following:
ESS: 95.27659 (ESS/N): 0.476383
Summary of weights generated, scaled to original weight:
(1 = no change, <1 = downweighted, >1 = upweighted)
      v1
Min.   :0.02081
1st Qu.:0.34025
Median :0.68355
Mean   :1.00000
3rd Qu.:1.18937
Max.   :5.53786
Original average outcomes by arm prior to weighting:
$control
[1] 0.7319588
$treatment
[1] 0.2038835
New average outcomes by arm after weighting:
$control
[1] 0.8393019
$treatment
[1] 0.1733996
See $results for:
[1] "average_outcome_group1" "average_outcome_group2" "match_adjusted_means"
[4] "new_average_outcome_group1" "new_average_outcome_group2"
```

In the previous output, the summary binary outcome is indicated for the control and treatment arms both before and after matching. These show that patients in the control arm on average experienced the outcome more frequently and that after matching, this difference from the intervention was more acute. If the relative treatment effect information between the comparator and anchor treatment is known and can be entered, then the `maic()` function may be used for a full Matching-Adjusted Indirect Comparison. For the launch version of `maicer`, only binary outcomes may be processed by the function to carry out the appropriate indirect comparison. The `outcome_type` argument is set to “binary” by default. Here a relative treatment effect of 3.02 with standard error 0.1036094071 has been entered:

```

Console Jobs x
~/IRC Research M.Sc. (Ind.)/Tool Development/maicer/ ↗
> anmaic <- maic(some_IPD, covariates_to_match, outcome_identifier = "y", binary_event_type = "positive", arm_iden
ntifier = "Group", outcome_type = "binary", comparator_to_anchor = 3.02, se_comparator_to_anchor = 0.1036094071)
> anmaic
A Matching-Adjusted Indirect Comparison resulted in the following:

Population-Adjusted Relative Treatment effect: 2.764455
Standard error: 0.1801516

The MAIC had an Effective Sample Size (ESS) of: 95.62413
ESS as a proportion of sample size: 0.4781206

See summary() for more information. If comparator to anchor trial information has been entered, see plot() for a
forest plot of adjusted vs unadjusted results.
> |

```

The `maic()` function also takes optional user input to name the intervention, anchor and comparator treatments. The full table of results is returned as part of the `summary()` function:

```

Console Jobs x
~/IRC Research M.Sc. (Ind.)/Tool Development/maicer/ ↗
> named_maic <- maic(some_IPD, covariates_to_match, "y", binary_event_type = "positive", arm_identifier = "Grou
p", outcome_type = "binary", comparator_to_anchor = 3.02, se_comparator_to_anchor = 0.1036094071, name_interventi
on = "loloplitin", name_anchor = "sonaliptin", name_comparator = "trizepibronate")
> summary(named_maic)
A Matching-Adjusted Indirect Comparison resulted in the following:

Population-Adjusted Relative Treatment effect: 2.764455
Standard error: 0.1801516

The MAIC had an Effective Sample Size (ESS) of: 95.62413
ESS as a proportion of sample size: 0.4781206

The overall results:# A tibble: 5 x 7
  id Comparison Estimate var lo hi type
<int> <chr> <dbl> <dbl> <dbl> <dbl> <chr>
1 1 "loloplitin \nvs\n sonaliptin" 2.76 0.180 1.93 3.60 MAIC
2 2 "loloplitin \nvs\n sonaliptin" 1.89 0.101 1.27 2.51 Unadjusted
3 3 "trizepibronate \nvs\n sonaliptin" 3.02 0.104 2.39 3.65 Unadjusted
4 4 "trizepibronate \nvs\n loloplitin" 0.256 0.284 -0.789 1.30 MAIC
5 5 "trizepibronate \nvs\n loloplitin" 1.13 0.205 0.243 2.02 Unadjusted

[1] "arms" "average_outcome"
[3] "beta" "centred"
[5] "comparator_to_anchor" "ess"
[7] "essprop" "match_adjusted_covariates"
[9] "matching_set" "outcomes"
[11] "population_adjusted_effect" "population_adjusted_se_estimate"
[13] "relative_adjusted_effect" "relative_adjusted_effect_se"
[15] "relative_unadjusted_effect" "relative_unadjusted_effect_se"
[17] "rescaled" "rescaled_weights_hist"
[19] "results" "se_comparator_to_anchor"
[21] "targets" "unadjusted_intervention_to_anchor"
[23] "unadjusted_intervention_to_anchor_se" "verification"
[25] "verification_plot" "weights"
> |

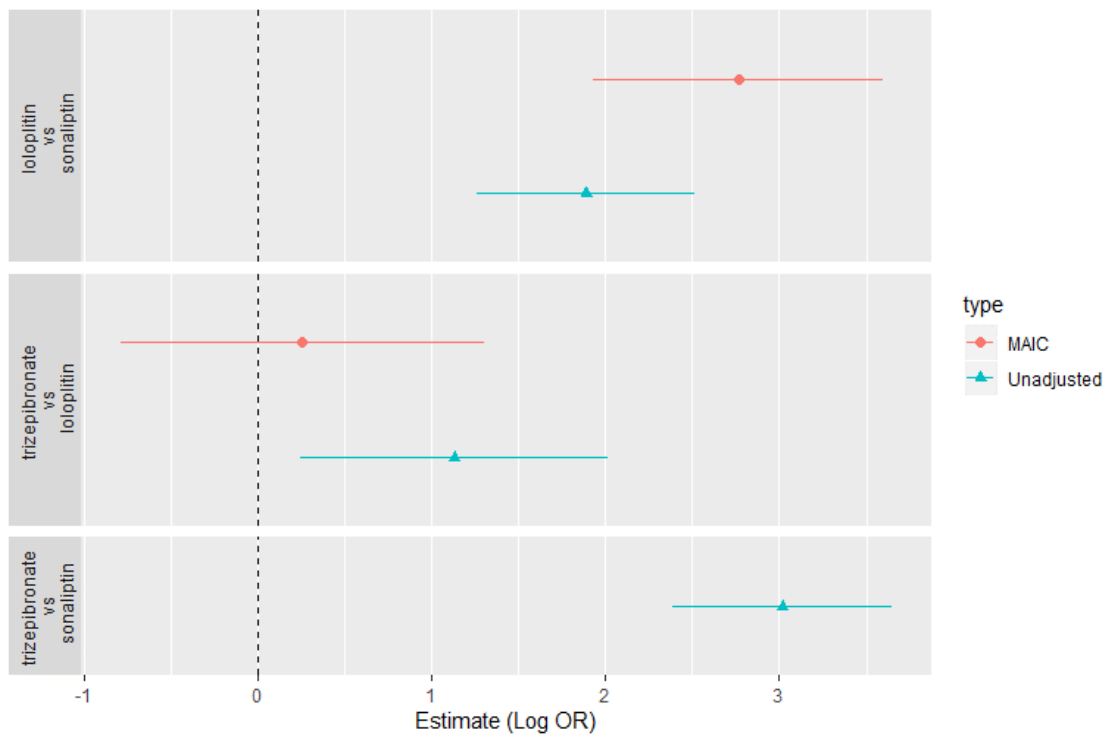
```

This output indicates that the relative treatment effect between the intervention and comparator treatments is 2.764 on the log odds scale (compared with 1.89 prior to matching-adjustment). In this instance it was understood that experiencing the event is positive for the patient; information which was passed into the `maic()` function via `binary_event_type = "positive"`. When treatment X is being compared against treatment Y, a value greater than 0 indicates patients administered treatment X experience the event more often and hence favours treatment X. A value less than 0 favours treatment Y. The user labelled the intervention treatment as "loloplitin", the anchor treatment as "sonaliptin" and the comparator treatment as "trizepibronate". The main

comparison of interest is therefore loloplitin vs trizepibronate. The results before and after matching suggest that loloplitin is superior to sonalipitin and these results are statistically significant. The results before matching then suggest that trizepibronate is superior to loloplitin and this result is statistically significant. However, the MAIC estimate suggests that while trizepibronate may be superior to loloplitin, the difference is far less acute and is not statistically significant.

In this example no true values were provided to show how biased each estimate was. It is important to recognise that MAIC does not necessarily reduce bias from true values in every case, although this is naturally a key objective.

Finally, using the `plot()` function will return a forest plot of the results for the various comparisons:



Noting that the event is positive for the health of the patient, for the plot above, in a comparison of treatment X vs treatment Y, values to the right of the vertical dashed line (representing log odds ratio of 0) favour treatment X. Values positioned to the left of 0 favour treatment Y (as treatment X patients are observed to be experiencing the event less frequently).

In the forest plot above, the MAIC estimates of relative treatment effect have been plotted alongside the standard Bucher comparisons which have not adjusted for the trial population. In the anchored MAIC analysis, it is assumed that the comparator trial results hold and they are not adjusted. It is the adjusted results of the intervention vs anchor treatment trial which affect the final relative treatment effect between the intervention and comparator of interest.

The MAIC estimates attract a larger standard error than the standard Bucher comparisons. The difference between the unadjusted and MAIC estimates suggests that after accounting for population differences, the true relative treatment effect between loloplitin and trizepibronate is less acute than was the case using unadjusted estimates of treatment for the indirect comparison. The results are also no longer statistically significant and so the MAIC estimate would suggest that these comparative results between loloplitin and trizepibronate are inconclusive.

Chapter 5 Conclusions

In the field of HTA and drug reimbursement, there is a clear demand from drug companies to make optimal use of evidence synthesis methods including population adjustment and MAIC in order to provide the most relevant detail for a drug reimbursement decision. Concurrently, HTA agencies seek to obtain the most accurate picture of the true effectiveness of a pharmaceutical product relative to other alternatives. They will take into consideration the level of uncertainty inherent in evidence not derived from head-to-head, double-blinded randomised controlled trials. The literature review in this thesis has shown that MAIC is becoming more frequently adopted within comparative effectiveness research; across multiple therapeutic areas and in a number of circumstances. While there does not appear to be any sign that this technique is falling out of use, sufficient guidance on its implementation remains lacking. The TSD provided a firm grounding for best practice on how the technique should be applied in HTA submissions to NICE. However, the literature review of published MAICs would suggest that implementation of the technique is non-standardised and is subject to a degree of variation as a result. In the worst case, the current literature on MAIC interprets the technique as little beyond simply “matching variables”.

The work from this thesis offers an attempt to standardise the implementation of the technique through means of a package in R. We believe this, and extending the work further, as suggested in Section 5.1, offers the best solution to consolidating the detail on the technique as well as providing investigators, reviewers and other stakeholders with the means to reproduce analyses. This can give a greater degree of transparency and assurance in the results derived from the method and facilitates comparative effectiveness research. We believe the choice of R is appropriate for the reasons outlined in Chapter 4.

5.1 Further Work

The work embodied in this research has motivated a number of possible extensions to the development of the R package. Further research into the fundamental properties of the MAIC methodology through simulation studies is recommended by the TSD. The development of this package offers a vehicle for conducting such studies, for example, changes to ESS due to varying patient population overlaps. Some suggestions for extensions to this work follow.

Consideration for all other types of outcomes would be of huge benefit to the development of the R package. As an alternative solution, the package aims to return the user with the weights generated by the MAIC algorithm so that they may be used to weight patient outcomes, whereupon the analyst may handle the data as required. Currently, there exists an application published by the Canadian HTA agency (CADTH) which assists in performing the calculations required for indirect treatment comparisons.

The package in its prototype form does not consider multilevel covariates for matching. Theoretically, this could be implemented using a design matrix to expand on the multilevel covariate into a series of dummy variables; each describing the various levels of the covariate. These would then be treated as binary covariates and passed into the matching set. In Section 4.4.2, the challenges for adapting the package to match IPD to aggregate median statistics were presented. Further extensions could involve accommodating the less usual summary data found in published clinical trials and MAIC analyses such as quartiles, correlations and higher moments (although in the case of standard deviation, the second moment is necessarily matched).

Stemming from the review of the MAIC methodology and its variations, one possible extension of the R package would be to allow for other weighting procedures to be incorporated in the tool such that the user may select the method they wish to implement. This would be less preferable to the other suggested extensions, since the premise of this work was to articulate the MAIC methodology in a standardised, reproducible and coherent manner. Allowing for alternative weighting methodologies as described in (Petto et al. 2018) or possibly as in (Regnier et al. 2016), carries the risk of somewhat taking away from this effort. It may also obscure the impact of MAIC. However, as a research exercise it could be interesting to view the properties of other weighting methods and their implications for HTA.

To encourage uptake of the R package, a user-friendly interface could be developed using R Shiny which would overlay the package functions. Assuming an adequately intuitive interface is developed, the tool could serve a wider audience who may be interested in implementing MAIC. In designing an interface, it may be necessary to restructure the code. An important consideration for a potential R Shiny app would be whether to host it on a server or locally. While hosting the app on a public website could be useful for raising awareness of the existence of the package, potential users may not have the freedom to upload their IPD via the online interface. IPD is likely to require secure storage and uploading such data online raises questions of security and data protection. With access to the R Shiny app script files, it would be possible to run the app internally using RStudio, but this negates the benefits of not requiring R and R Shiny pre-installed on the user's computer.

The focus of the R package implementation was on the anchored scenario. The unanchored scenario is relatively straightforward to carry out and the weights can be obtained using the `match_adjustment` object. Reorganising the code to draw a more explicit distinction between the two scenarios could be useful for the structure of the R package. This could involve extending the `match_adjustment` object, for example.

A final potential area for development of the package is to introduce greater automation and generality of the code with respect to handling user input. As an example, the user requires their

IPD to follow a number of standards (although the template is not highly stringent), such as formatting of dichotomous variables to be coded to {0,1}.

5.2 Final Remarks

This thesis has confirmed that the full implications for use of MAIC are not fully understood. In summary:

- It is unclear at what threshold it is appropriate to perform MAIC with respect to varying degrees of differing patient populations and the implication for modelling practice.
- It is unclear to what extent variations on the methodology for calculating weights as well as other methodological variations impact on the results of analyses.
- In the absence of IPD on a comparator treatment and use of AgD for matching purposes, it is unclear how matching to a median statistic can be best implemented.
- Further research into the appropriateness of matching by method of moments is warranted.

In order to adopt routine appropriate usage of MAIC, it is necessary to standardise its implementation so that results can be rendered comparable across technology appraisals. The work from this thesis has sought to identify a standard approach to applying MAIC in addition to method variations. It has identified the risk of disparity that currently exists arising from the multitude of variations in the methodology. As a solution, a template for implementing the method has been created through the R package. While multiple extensions to this tool are possible and should be pursued, the principal goal is that the effort of standardising MAIC continues. The exact circumstances in which it is appropriate to apply MAIC have been touched upon in this thesis but fall outside the focus of this project. Assuming appropriate usage, standardised implementation of MAIC can help to ensure optimal healthcare decision-making.

Appendix A Binomial Distribution as a Member of the Exponential Family of Distributions

From (Dobson and Barnett 2008), it can be shown that the Binomial distribution is a member of the Exponential Family of Distributions. To show this, it is necessary to show that its probability density function can be written in the form:

$$f(y; \theta) = s(y)t(\theta)e^{a(y)b(\theta)}$$

Where y describes a single random variable whose probability distribution depends on a single parameter θ .

This equation can be rewritten so that $s(y) = \exp d(y)$ and $t(\theta) = \exp c(\theta)$ as follows:

$$f(y; \theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

From Equation 2, the probability density function for the Binomial distribution is:

$$f(y; \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y} \quad (2)$$

Or where $n = 1$,

$$f(y; \theta) = \theta^y (1 - \theta)^{1-y}$$

Applying the exponential function:

$$\exp \left[\binom{n}{y} \theta^y (1 - \theta)^{1-y} \right]$$

Taking logs of the exponent:

$$\begin{aligned} & \exp \left[\binom{n}{y} \theta^y (1 - \theta)^{1-y} \right] \\ &= \exp \left[\log \binom{n}{y} \log(\theta^y) \log((1 - \theta)^{1-y}) \right] \\ &= \exp \left[\log \binom{n}{y} + \log(\theta^y) + \log((1 - \theta)^{1-y}) \right] \\ &= \exp \left[\log \binom{n}{y} + y \log(\theta) + (1 - y) \log(1 - \theta) \right] \\ &= \exp \left[\log \binom{n}{y} + y \log(\theta) + \log(1 - \theta) - y \log(1 - \theta) \right] \\ &= \exp \left[y(\log(\theta) - \log(1 - \theta)) + \log(1 - \theta) + \log \binom{n}{y} \right] \end{aligned}$$

$$= \exp \left[y \log \left(\frac{\theta}{1-\theta} \right) + \log(1-\theta) + \log \binom{n}{y} \right]$$

Therefore $a(y)b(\theta) = y \log \left(\frac{\theta}{1-\theta} \right)$ and $c(\theta) = \log(1-\theta)$ and $d(y) = \log \binom{n}{y}$.

Appendix B List of Reviewed MAIC Publications

The following publications were reviewed and are listed below in no particular order:

- (Signorovitch et al. 2011): vildagliptin (IPD) vs sitagliptin (AgD)
- (Regnier et al. 2016): ranibizumab (IPD) vs aflibercept (AgD)
- (Van Sanden et al. 2018): (Overall Survival outcome, unanchored MAIC) daratumumab (IPD) vs pomalidomide + dexamethasone (AgD)
- (Van Sanden et al. 2016): simeprevir + peginterferon alfa 2a + ribavirin (IPD) vs peginterferon alfa 2a + ribavirin (AgD)
- (Odom et al. 2017): sonidegib (IPD) vs vismodegib (AgD)
- (Warren et al. 2018): ixekizumab (IPD) vs secukinumab (AgD)
- (Signorovitch et al. 2013): everolimus (IPD) vs sunitinib (AgD)
- (Ishak et al. 2018): sunitinib (IPD) vs everolimus (AgD)
- (Majer et al. 2017): panobinostat + borezomib + dexamethasone (IPD) vs lenalidomide + dexamethasone (IPD*) and panobinostat + borezomib + dexamethasone (IPD) vs pomalidomide + dexamethasone (IPD*)
- (Tremblay et al. 2018): ribociclib + letrozole (IPD) vs palbociclib + letrozole (AgD)
- (Song et al. 2019): blinatumomab (IPD) vs inotuzumab ozogamicin (AgD)
- (Atkins et al. 2019): nivolumab + ipilimumab (IPD) vs dabrafenib + trametinib (AgD) and nivolumab + ipilimumab (IPD) vs vemurafenib + cobimetinib (AgD)
- (Proskorovsky et al. 2018): axitinib (IPD) vs cabozantinib (AgD) and axitinib (IPD) vs everolimus (AgD)
- (Signorovitch et al. 2015) nilotinib (IPD) vs dasatinib (AgD)
- (Maksymowych et al. 2018) secukinumab (IPD) vs adalimumab (AgD)
- (Signorovitch et al. 2010) adalimumab (IPD) vs etanercept (AgD)

*IPD was approximated using the algorithm by (Guyot et al. 2012) in order to draw a comparison in terms of Progression-Free Survival (PFS) and Overall Survival (OS) but was not weighted.

Bibliography

- Atkins, M. B., Tarhini, A., Rael, M., Gupte-Singh, K., O'Brien, E., Ritchings, C., Rao, S. and McDermott, D. F. (2019) Comparative efficacy of combination immunotherapy and targeted therapy in the treatment of BRAF-mutant advanced melanoma: a matching-adjusted indirect comparison. *Immunotherapy*, 11(7), pp. 617-629.
- Berlin, J. A., Santanna, J., Schmid, C. H., Szczech, L. A. and Feldman, H. I. (2002) Individual patient-versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Statistics in medicine*, 21(3), pp. 371-387.
- Briggs, A., Sculpher, M. and Claxton, K. (2006) *Decision modelling for health economic evaluation*, Oxford University Press.
- Broyden, C. G. (1970) The convergence of a class of double-rank minimization algorithms: 2. The new algorithm. *IMA journal of applied mathematics*, 6(3), pp. 222-231.
- Bucher, H. C., Guyatt, G. H., Griffith, L. E. and Walter, S. D. (1997) The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of clinical epidemiology*, 50(6), pp. 683-691.
- Carlin, A. M., Zeni, T. M., English, W. J., Hawasli, A. A., Genaw, J. A., Krause, K. R., Schram, J. L., Kole, K. L., Finks, J. F. and Birkmeyer, J. D. (2013) The comparative effectiveness of sleeve gastrectomy, gastric bypass, and adjustable gastric banding procedures for the treatment of morbid obesity. *Annals of surgery*, 257(5), pp. 791-797.
- DerSimonian, R. and Laird, N. (1986) Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3), pp. 177-188.
- Dias, S., Sutton, A. J., Ades, A. and Welton, N. J. (2013) Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making*, 33(5), pp. 607-617.
- Dias, S., Sutton, A. J. and Ades, A. E. (2011a) *NICE DSU Technical Support Document 3: Heterogeneity: subgroups, meta-regression, bias and bias-adjustment*.
- Dias, S., Welton, N. J., Sutton, A. J. and Ades, A. E. (2011b) *NICE DSU Technical Support Document 2: A Generalised Linear Modelling Framework for Pairwise and Network Meta-Analysis of Randomised Controlled Trials*.
- Dobson, A. J. and Barnett, A. G. (2008) *An introduction to generalized linear models*, *Texts in Statistical Science*, Third ed.
- Donegan, S., Williamson, P., D'Alessandro, U., Garner, P. and Smith, C. T. (2013) Combining individual patient data and aggregate data in mixed treatment comparison meta-analysis: individual patient data may be beneficial if only for a subset of trials. *Statistics in medicine*, 32(6), pp. 914-930.

- Eichler, H.-G., Bloechl-Daum, B., Abadie, E., Barnett, D., König, F. and Pearson, S. (2010) Relative efficacy of drugs: an emerging issue between regulatory agencies and third-party payers. *Nature reviews Drug discovery*, 9(4), pp. 277.
- EUNetHTA (2013) *Comparators & Comparisons: Direct and indirect comparisons*. WP5: EUNetHTA.
- Filipovic-Pierucci, A., Zarca, K. and Durand-Zaleski, I. (2017) *Markov Models for Health Economic Evaluation: The R Package heemod*, Available: <https://pierucci.org/heemod>.
- Fletcher, R. (1970) A new approach to variable metric algorithms. *The computer journal*, 13(3), pp. 317-322.
- GitHub (2019) Available: <http://github.com/> [Accessed 25th September 2019].
- Goldfarb, D. (1970) A family of variable-metric methods derived by variational means. *Mathematics of computation*, 24(109), pp. 23-26.
- Greenland, S. and Morgenstern, H. (1989) Ecological bias, confounding, and effect modification. *International journal of epidemiology*, 18(1), pp. 269-274.
- Guyot, P., Ades, A., Ouwens, M. J. and Welton, N. J. (2012) Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC medical research methodology*, 12(1), pp. 9.
- Hawkins, N., Scott, D. A. and Woods, B. (2009) How far do you go? Efficient searching for indirect evidence. *Medical Decision Making*, 29(3), pp. 273-281.
- Higgins, J. and Green, S. (2011) *Cochrane Handbook for Systematic Reviews of Interventions*, [online], available: www.handbook.cochrane.org [accessed 26th September 2019].
- Hoaglin, D. C., Hawkins, N., Jansen, J. P., Scott, D. A., Itzler, R., Cappelleri, J. C., Boersma, C., Thompson, D., Larholt, K. M. and Diaz, M. (2011) Conducting indirect-treatment-comparison and network-meta-analysis studies: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 2. *Value in Health*, 14(4), pp. 429-437.
- Hosmer, D. W. and Lemeshow, S. (2000) *Applied Logistic Regression*, Second ed.
- Immunotherapy (2019) Available: <https://www.futuremedicine.com/journals/imt/aims> [Accessed 26th September 2019].
- Ishak, K. J., Proskorovsky, I. and Benedict, A. (2015) Simulation and Matching-Based Approaches for Indirect Comparison of Treatments. *Pharmacoeconomics*, 33(6), pp. 537-549.

- Ishak, K. J., Rael, M., Hicks, M., Mittal, S., Eatock, M. and Valle, J. W. (2018) Relative effectiveness of sunitinib versus everolimus in advanced pancreatic neuroendocrine tumors: an updated matching-adjusted indirect comparison. *Journal of comparative effectiveness research*, 7(10), pp. 947-958.
- Jansen, J. P. (2012) Network meta-analysis of individual and aggregate level data. *Research Synthesis Methods*, 3(2), pp. 177-190.
- Jansen, J. P., Fleurence, R., Devine, B., Itzler, R., Barrett, A., Hawkins, N., Lee, K., Boersma, C., Annemans, L. and Cappelleri, J. C. (2011) Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 1. *Value in Health*, 14(4), pp. 417-428.
- Kamangar, F. (2012) Effect modification in epidemiology and medicine. *Archives of Iranian Medicine (AIM)*, 15(9).
- Leisch, F. (2008) Creating R packages: A Tutorial. in Brito, P., (ed.) *Proceedings in Computational Statistics*, Heidelberg, Germany.
- Lumley, T. (2002) Network meta-analysis for indirect treatment comparisons. *Statistics in medicine*, 21(16), pp. 2313-2324.
- Majer, I., van de Wetering, G., Polanyi, Z., Krishna, A., Gray, E. and Roy, A. (2017) Panobinostat plus bortezomib versus lenalidomide in patients with relapsed and/or refractory multiple myeloma: a matching-adjusted indirect treatment comparison of survival outcomes using patient-level data. *Applied health economics and health policy*, 15(1), pp. 45-55.
- Maksymowych, W. P., Strand, V., Nash, P., Yazici, Y., Thom, H., Hunger, M., Kalyvas, C., Gandhi, K. K., Porter, B. and Jugl, S. M. (2018) Comparative effectiveness of secukinumab and adalimumab in ankylosing spondylitis as assessed by matching-adjusted indirect comparison. *European journal of rheumatology*, 5(4), pp. 216.
- Malangone, E. and Sherman, S. (2011) Matching-Adjusted Indirect Comparison Analysis Using Common SAS® 9.2 PROCEDURES. In Paper Presented to the SAS Global Forum 2011 Conference, Cary, NC, USA.
- METACRAN (2019) Available: <https://www.r-pkg.org/> [Accessed 4th February 2019].
- National Institute for Health and Care Excellence (2013) *Guide to the methods of technology appraisal 2013*. London.
- Neupane, B., Richer, D., Bonner, A. J., Kibret, T. and Beyene, J. (2014) Network meta-analysis using R: a review of currently available automated packages. *PloS one*, 9(12), pp. e115065.

- Norton, E. C., Miller, M. M., Wang, J. J., Coyne, K. and Kleinman, L. C. (2012) Rank reversal in indirect comparisons. *Value in Health*, 15(8), pp. 1137-1140.
- Odom, D., Mladsi, D., Purser, M., Kaye, J. A., Palaka, E., Charter, A., Jensen, J. A. and Sellami, D. (2017) A matching-adjusted indirect comparison of sonidegib and vismodegib in advanced basal cell carcinoma. *Journal of skin cancer*, 2017.
- Olmos, A. and Govindasamy, P. (2015) Propensity scores: a practical introduction using R. *Journal of MultiDisciplinary Evaluation*, 11(25), pp. 68-88.
- Petto, H., Kadziola, Z., Brnabic, A., Saure, D. and Belger, M. (2018) Alternative Weighting Approaches for Anchored Matching-Adjusted Indirect Comparisons (MAICs) via a Common Comparator. *Value in Health*.
- Phillippo, D., Ades, T., Dias, S., Palmer, S., Abrams, K. and Welton, N. (2017) Population-adjusted treatment comparisons: estimates based on MAIC (Matching-Adjusted Indirect Comparisons) and STC (Simulated Treatment Comparisons). in *ISPOR 22nd Annual International Meeting*.
- Phillippo, D., Ades, T., Dias, S., Palmer, S., Abrams, K. R. and Welton, N. (2016) NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submissions to NICE.
- Phillippo, D., Dias, S., Elsada, A., Ades, A. and Welton, N. (2019) Population Adjustment Methods for Indirect Comparisons: A Review of National Institute for Health and Care Excellence Technology Appraisals. *International Journal of Technology Assessment in Health Care*.
- Phillippo, D. M., Ades, A. E., Dias, S., Palmer, S., Abrams, K. R. and Welton, N. J. (2018) Methods for Population-Adjusted Indirect Comparisons in Health Technology Appraisal. *Medical Decision Making*, 38(2), pp. 200-211.
- Proskorovsky, I., Benedict, A., Negrier, S., Bargo, D., Sandin, R., Ramaswamy, K., Desai, J., Cappelleri, J. C. and Larkin, J. (2018) Axitinib, cabozantinib, or everolimus in the treatment of prior sunitinib-treated patients with metastatic renal cell carcinoma: results of matching-adjusted indirect comparison analyses. *BMC cancer*, 18(1), pp. 1271.
- R Core Team (2018) R: A Language and Environment for Statistical Computing. in, Vienna, Austria: R Foundation for Statistical Computing.
- R Core Team (2019) *Writing R Extensions*, Available: <https://cran.r-project.org/doc/manuals/r-release/R-exts.html> [Accessed 17th September 2019].
- Regnier, S. A., Alsop, J., Wright, J., Nixon, R., Staines, H. and Fajnkuchen, F. (2016) Review and comparison of methodologies for indirect comparison of clinical trial results: an illustration with ranibizumab and aflibercept. *Expert review of pharmacoeconomics & outcomes research*, 16(6), pp. 793-801.

- Riley, R. D., Lambert, P. C. and Abo-Zaid, G. (2010) Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*, 340, pp. c221.
- Rosenbaum, P. R. (1987) Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398), pp. 387-394.
- Rosenbaum, P. R. and Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), pp. 41-55.
- Sekhon, J. S. (2011) Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R. *Journal of Statistical Software*, 42(7), pp. 52.
- Shanno, D. F. (1970) Conditioning of quasi-Newton methods for function minimization. *Mathematics of computation*, 24(111), pp. 647-656.
- Signorovitch, J., Swallow, E., Kantor, E., Wang, X., Klimovsky, J., Haas, T., Devine, B. and Metrakos, P. (2013) Everolimus and sunitinib for advanced pancreatic neuroendocrine tumors: a matching-adjusted indirect comparison. *Experimental hematology & oncology*, 2(1), pp. 32.
- Signorovitch, J. E., Betts, K. A., Reichmann, W. M., Thomason, D., Galebach, P., Wu, E. Q., Chen, L. and DeAngelo, D. J. (2015) One-year and long-term molecular response to nilotinib and dasatinib for newly diagnosed chronic myeloid leukemia: a matching-adjusted indirect comparison. *Current medical research and opinion*, 31(2), pp. 315-322.
- Signorovitch, J. E., Sikirica, V., Erder, M. H., Xie, J., Lu, M., Hodgkins, P. S., Betts, K. A. and Wu, E. Q. (2012) Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research. *Value in Health*, 15(6), pp. 940-947.
- Signorovitch, J. E., Wu, E. Q., Andrew, P. Y., Gerrits, C. M., Kantor, E., Bao, Y., Gupta, S. R. and Mulani, P. M. (2010) Comparative effectiveness without head-to-head trials. *Pharmacoeconomics*, 28(10), pp. 935-945.
- Signorovitch, J. E., Wu, E. Q., Swallow, E., Kantor, E., Fan, L. and Gruenberger, J.-B. (2011) Comparative Efficacy of Vildagliptin and Sitagliptin in Japanese Patients with Type 2 Diabetes Mellitus. *Clinical Drug Investigation*, 31(9), pp. 665-674.
- Sikirica, V., Findling, R. L., Signorovitch, J., Erder, M. H., Dammerman, R., Hodgkins, P., Lu, M., Xie, J. and Wu, E. Q. (2013) Comparative efficacy of guanfacine extended release versus atomoxetine for the treatment of attention-deficit/hyperactivity disorder in children and adolescents: applying matching-adjusted indirect comparison methodology. *CNS drugs*, 27(11), pp. 943-953.
- Sivaprasad, S., Regnier, S. A., Fajnkuchen, F., Wright, J., Berger, A. R., Mitchell, P. and Larsen, M. (2016) Using Patient-Level Data to Develop Meaningful Cross-Trial Comparisons of Visual Impairment in Individuals with Diabetic Macular Edema. *Advances in therapy*, 33(4), pp. 597-609.

- Song, J., Ma, Q., Gao, W., Cong, Z., Xie, J., Zimmerman, Z., Belton, L., Franklin, J. and Palmer, S. (2019) Matching-Adjusted Indirect Comparison of Blinatumomab vs. Inotuzumab Ozogamicin for Adults with Relapsed/Refractory Acute Lymphoblastic Leukemia. *Advances in therapy*, pp. 1-12.
- Stewart, L. A. and Tierney, J. F. (2002) To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the health professions*, 25(1), pp. 76-97.
- Sutton, A., Ades, A. E., Cooper, N. and Abrams, K. (2008) Use of Indirect and Mixed Treatment Comparisons for Technology Assessment. *Pharmacoeconomics*, 26(9), pp. 753-767.
- Tierney, N. (2016) A Simple Guide to S3 Methods. *arXiv preprint arXiv:1608.07161*.
- Tremblay, G., Chandiwana, D., Dolph, M., Hearnden, J., Forsythe, A. and Monaco, M. (2018) Matching-adjusted indirect treatment comparison of ribociclib and palbociclib in HR+, HER2- advanced breast cancer. *Cancer management and research*, 10, pp. 1319.
- Van Sanden, S., Ito, T., Diels, J., Vogel, M., Belch, A. and Oriol, A. (2018) Comparative Efficacy of Daratumumab Monotherapy and Pomalidomide Plus Low-Dose Dexamethasone in the Treatment of Multiple Myeloma: A Matching Adjusted Indirect Comparison. *The oncologist*, 23(3), pp. 279-287.
- Van Sanden, S., Pisini, M., Duchesne, I., Mehnert, A. and Belsey, J. (2016) Indirect comparison of the antiviral efficacy of peginterferon alpha 2a plus ribavirin used with or without simeprevir in genotype 4 hepatitis C virus infection, where common comparator study arms are lacking: a special application of the matching adjusted indirect comparison methodology. *Current medical research and opinion*, 32(1), pp. 147-154.
- Veroniki, A. A., Straus, S. E., Soobiah, C., Elliott, M. J. and Tricco, A. C. (2016) A scoping review of indirect comparison methods and applications using individual patient data. *BMC medical research methodology*, 16(1), pp. 47.
- Wang, J., Odom, D., Chirila, C. and Zheng, Q. (2015) Evaluation Of Matching-Adjusted Indirect Comparison Implemented By A Resampling Method. in *ISPOR 20th Annual International Meeting*, Philadelphia, PA, USA.
- Warren, R., Brnabic, A., Saure, D., Langley, R., See, K., Wu, J., Schacht, A., Mallbris, L. and Nast, A. (2018) Matching-adjusted indirect comparison of efficacy in patients with moderate-to-severe plaque psoriasis treated with ixekizumab vs. secukinumab. *British Journal of Dermatology*, 178(5), pp. 1064-1071.
- Wells, G. A., Sultan, S. A., Chen, L., Khan, M. and Coyle, D. (2009) *Indirect Evidence: Indirect Treatment Comparisons in Meta-Analysis*. Ottawa: Canadian Agency for Drugs and Technologies in Health (CADTH).

White, H. (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *econometrica*, 48(4), pp. 817-838.

Wickham, H. (2015) *R Packages*, First ed., O'Reilly Media, Inc.